Facultad de Ciencias Exactas
Universidad Nacional del Centro de la Provincia de Buenos Aires

Machine learning for ophthalmic screening and diagnostics from fundus images

por
José Ignacio Orlando

Thesis submitted in partial fulfillment of the requirements for the degree of
Doctor en Matemática Computacional e Industrial

Advisors:
Dr. Matthew B. Blaschko
Dr. Mariana del Fresno

2017

Resumen

Las fotografías de fondo de ojo representan una técnica no-invasiva muy empleada por oftalmólogos para detectar retinopatía diabética (RD) y glaucoma, dos de las principales causas mundiales de ceguera prevenible.

En esta tesis se presentan algunas contribuciones al análisis automático de estas imágenes mediante técnicas de aprendizaje automático.

Inicialmente se propone un método de segmentación de vasos sanguíneos basado en el aprendizaje de campos condicionales aleatorios totalmente conectados mediante máquinas de vectores de soporte de salida estructurada. Este enfoque permite obtener representaciones precisas del árbol vascular, que son luego utilizadas en el contexto de dos técnicas para la detección de glaucoma y RD.

Para detectar glaucoma, se plantea transferir redes neuronales convolucionales (CNNs) preentrenadas con datos no médicos. Las imágenes son adaptadas inicialmente mediante métodos de preprocesamiento tradicionales, y los descriptores extraídos se utilizan para entrenar modelos regularizados de regresión logística, obteniendo resultados competitivos con el estado del arte.

Finalmente, se introduce una técnica para detectar lesiones rojas típicas de la RD basada en hibridar descriptores aprendidos utilizando una CNN y otros diseados manualmente. Posteriormente se utiliza un Bosque Aleatorio entrenado con estos valores para identificar lesiones candidatas, con alta eficacia en diversas bases de datos.

Abstract

Fundus images are a non-invasive imaging modality that is typically used by ophthalmologists to assess the retina. They are widely applied for detecting diabetic retinopathy (DR) and glaucoma, which are leading causes of preventable blindness in the world.

In this thesis we contribute with novel tools for automated fundus image analysis based on machine learning. First, we propose a vessel segmentation method based on learning fully connected conditional random fields using structured output support vector machines. This approach allows to recover accurate segmentations of the retinal vasculature that are afterwards applied in the context of two deep learning based techniques for glaucoma and DR detection.

For automated glaucoma assessment, we propose to transfer convolutional neural networks (CNNs) that were pre-trained using non-medical data. Images are adapted with state of the art preprocessing methods before feeding the CNNs. The extracted features are used to train regularized logistic regression classifiers, achieving results that are competitive with other methods.

For DR screening, we introduce a red lesion detection approach based on hybridizing deep learned and hand crafted features. A Random Forest classifier is trained on these features to identify true lesion candidates, reporting state of the art performance in benchmark data sets.

Acknowledgements

I never though that writing the acknowledgments of a thesis would be that difficult. It turns out to be even harder when one tries by all means not to fall into clichs, or to thank to all who collaborate with me to reach the end of this long road. "Long road", well, first clich. This will be tough.

First of all, I want to thank my supervisors, Mariana del Fresno and Matthew B. Blaschko, for accompanying me and advising me during the development of this thesis. Mariana, I I greatly value your trust in letting me explore my vocation as a scientist. Matthew, your support both from a technical and a human point of view was crucial to me, and it is largely responsible for my success on this day. I learned a lot working with you, and I will always be grateful for it.

I would also like to thank the members of the jury, Dr. Virginia Ballarn (UNMdP), Dr. Diego Milone (UNR), Dr. Jos Massa (UNICEN) and Dr. Clarisa Snchez (Radbound University Medical Center, The Netherlands), for having spent part of his valuable time reading this thesis and proposing improvements and contributions to the final document.

I am also grateful to Alejandro Clausse, Pablo Lotito and Ignacio Larrabide, members of Pladema who trusted me and allowed me to work on other projects, derived from this thesis in some cases, and completely alien to it in others. I also want to thank Claudia Marinelli for helping me in the statistical analysis of my results, and Diego Dalponte, who trusted me in the task of teaching at the university.

I would also like to thank the clinicians and researchers who collaborated in the development of this doctoral thesis: Elena Prokofyeva, Karel van Keer and João Barbosa Breda. Their contributions from the clinical and epidemiological points of view were invaluable when devising the solutions proposed in this paper and posing

future challenges and applications.

I am also grateful to the educational and scientific institutions of Argentina, and to Nestor and Cristina Kirchner, presidents of Argentina that supported research. *No fue magia.*

To my friends, Mante, Emma, Mari, Cris, Nacha, Enzo, Adrin, Fernanda, Vicky and Mai. Just a big hug to them. Life with them around is always much happier.

I want to dedicate a few lines in these (long) acknowledgments to my colleagues in Pladema, especially to those who surrounded my headphones in the last few years. To Delfi, Lover, Mante, Laucha, Lean, Mati, Trapa, Blito, Zeque, Hernan, Sil and Nico, and also to Boro, to Mara, and (only a bit) to Javier, thanks for the *mate*, the gossips and the long Slack conversations. In the same way, a hug and thanks to the teammates I met in other parts of the world: to those in Paris, Siddhartha, Aline, Katherina, Eugenio and Puneet, and to those in Leuven, especially Maxim and Amal, two great friends in whose company the distance to Argentina always tends to 0.

Thanks also to all the wonderful people I met at the Argentinian House in Paris, who made me feel at home when I had an entire ocean between me and my family. Fede, Pilu, Maru, Allo and Gus, thanks for everything. And also thanks to Marcelo, former manager of the Argentinian House, who did the impossible so that everyone enjoyed the experience of living in the City of Light.

This is a clich, again, but I want to thank my family, of course. To my mum, dad and siblings, for accepting me as I am and for always supporting me when I need it. To my grandmother Norma, number one fan of everything that I do. To my aunts, cousins, to Nati and Martin, to all the giant family that I have in Las Flores.

And last but not least, to Sebas, the love of my life. Thanks for being always there.

To my grandmother, Norma. The one with the greatest heart in the world.
And to my grandparents, who are not here, but I'm pretty sure they are.

Contents

Acknowledgements **7**

1 Introduction **22**
1.1 Motivation 22
1.2 Machine learning 25
1.3 Contributions and thesis outline 26

2 Background **30**
2.1 The eye and the retina 30
2.2 Retinal diseases 33
2.2.1 Diabetic retinopathy 33
2.2.2 Glaucoma 36
2.3 Retinal imaging 39
2.3.1 Fundus photographs 39
2.3.2 Fluorescein angiography 42
2.3.3 Optical coherence tomography 43
2.4 Fundus image analysis for computer-assisted diagnosis 45
2.4.1 Overview 45
2.4.2 Vessel characterization approaches 46
2.4.3 Detection of pathological structures 49
2.4.4 Image characterization using global features 53
2.4.5 Deep learning based methods 55

3 Blood vessel segmentation **59**
3.1 Motivation 60

3.2 Methods . . . 63
3.2.1 Conditional Random Fields for vessel segmentation . . . 64
3.2.2 Learning CRFs with Structured Output SVM . . . 67
3.2.3 Features . . . 69
3.2.4 Scaling Models to Images of Different Resolution . . . 71
3.3 Materials and Evaluation . . . 73
3.3.1 Datasets . . . 73
3.3.2 Evaluation Metrics . . . 74
3.3.3 Model selection . . . 76
3.4 Experiments and Results . . . 77
3.4.1 Results . . . 78
3.4.2 Comparison with other methods . . . 84
3.5 Discussion . . . 89
3.6 Conclusions . . . 91

4 Transfer learning for glaucoma screening **92**
4.1 Motivation . . . 93
4.2 Methods . . . 95
4.2.1 Image preprocessing . . . 96
4.2.2 ℓ_1 and ℓ_2 regularized logistic regression . . . 98
4.3 Materials . . . 100
4.3.1 Convolutional neural networks . . . 100
4.3.2 Data sets . . . 100
4.4 Results . . . 101
4.5 Discussion . . . 104

5 Red lesion detection for DR screening **106**
5.1 Motivation . . . 107
5.2 Methods . . . 110
5.2.1 Candidate detection . . . 111
5.2.2 CNN-based features . . . 114
5.2.3 Hand-crafted feature extraction . . . 116
5.2.4 Candidate classification with Random Forest . . . 118

5.3 Experimental setup 121
5.3.1 Materials 121
5.3.2 Model selection 123
5.3.3 Evaluation metrics 124
5.4 Results 125
5.4.1 Per lesion evaluation 125
5.4.2 Per image evaluation 129
5.4.3 Feature assessment 131
5.5 Discussion 134
5.6 Conclusions 138

6 Conclusions and future lines of research **140**
6.1 Contributions 140
6.2 Future lines of research 142
6.2.1 Vessel segmentation and characterization 142
6.2.2 Automated glaucoma screening 143
6.2.3 Automated DR screening 145

Bibliography **146**

List of Figures

2.1 Schematic representation of the eye. Source: [42]. 31

2.2 Microvascular complications due to hyperglycemia. (a) Healthy microvessel. (b) Hyperglycemia weakens vessel walls and increments the blood flow, while lipids might be leaked. (c) Weaker vessel walls and the increase shear strees favours the formation of microaneurysms. (d) The increased pressure can eventually lead to the rupture of the microaneurysm and produce hemorrhages. 34

2.3 Diabetic retinopathy stages. Source: [19]. 35

2.4 Aqueous humor flow. In light blue, aqueous humor pathway. In red, iridocorneal angle. 36

2.5 Fundus image of a healthy subject, extracted from the HRF data set [145]. 39

2.6 Fundus camera. Source: Ralf Roletschek / Wikipedia [171]. 40

2.7 Volk iNview, a smartphone-based device for fundus image acquisition. Source: [148]. 40

2.8 Fundus images for retinal disease diagnosis. 41

2.9 Fluorescein angiography. Phases. Source: [167]. 43

2.10 OCT scan and retinal layers as obtained applying and automated segmentation method [112]. (a) X-Z image of the OCT volume. (b) Segmentation results, nerve fiber layer (NFL), ganglion cell layer (GCL), inner plexiform layer (IPL), inner nuclear layer (INL), outer plexiform layer (OPL), outer nuclear layer (ONL), inner segment outer segment junction (ISOSJ), outer segment photoreceptors (OPR), subretinal virtual space (SRVS-zero thikness in normals), and retinal pigment epithelium (RPE). Figure extracted from [191]. 44
2.11 Screenshot of the Singapore I Vessel Assessment (SIVA) tool. Source: [1]. 47
2.12 Examples of red lesions as observed in fundus images. Notice that some of them are barely noticeable. Source: DIARETDB1 training set [97]. 50
2.13 Examples of bright lesions as observed in fundus images. Source: DIARETDB1 training set [97]. 50
2.14 Retinal Nerve Fiber Layer (RNFL) defects due to glaucoma, as manifested in fundus photographs. Source: [75]. 52
2.15 Peripapillary atrophy due to glaucoma. Source: [117]. 52
2.16 Abnormality heat maps as obtained by the deep learning based method for DR detection proposed by Gargeya *et al.* [69]. Source: [69]. 57

3.1 A fundus photograph from the test set of DRIVE [141] and its vessel annotations as obtained by a human expert and the segmentation approach proposed in this thesis. 61
3.2 Image preprocessing and unary and pairwise features examples. (a) Original color image. (b) Inverted green band after border expansion. (c) Response to Nguyen *et al.* line detector ($l = 15$). (d) Response to Soares *et al.* 2D Gabor wavelet at the scale $a = 3$. (e) Inverted image after bias correction. (f) Image enhanced using Zana and Klein method ($l = 9$). 71
3.3 Segmentation results obtained on DRIVE. (a) Image 04 of DRIVE. (b) Ground truth labelling. (c) Segmentation obtained using only the unary potentials. (d) Segmentation obtained using the FC-CRF. . . . 80

3.4 Segmentation results obtained on CHASEDB1. (a) Image_05L of CHASEDB1. (b) Ground truth labelling. (c) Segmentation obtained using only the unary potentials. (d) Segmentation obtained using the FC-CRF. 81

3.5 Segmentation results obtained on a serious pathological case on STARE. (a) Image im0005. (b) First human observer annotations. (c) Second human observer annotations. (d) Segmentation obtained using only the unary potentials. (e) Segmentations obtained using the FC-CRF model. 82

3.6 Segmentation results obtained on HRF test set. (a) Image 11_g. (b) Manual annotation. (c) Segmentation obtained using only the unary potentials. (d) Segmentation obtained using the FC-CRF. 83

3.7 Example of narrow vessel detection under low contrast conditions. (a) Detail of Image 11_g. (b) Preprocessed image. (c) Manual annotation. (d) Segmentation obtained using only the unary potentials. (e) Segmentation obtained using the FC-CRF. 84

3.8 ROC curves on DRIVE, CHASEDB1 and HRF, using only the unary potentials (UP, slashed line) or the FC-CRF (solid line), and second human observer (HO) performance. 85

3.9 Computational cost of the FC-CRF inference in all the data sets used for evaluation. 88

4.1 Schematic representation of our method for transferring CNNs pre-trained from non-medical data to glaucoma detection in fundus images. 96

4.2 Preprocessing strategies evaluated. First group: without CLAHE. Second group: with CLAHE. From left to right: original image, cropped FOV, peripapillary area (PPA), and ONH. First row: original images. Second row: images after vessel inpainting. 97

4.3 Vessel subtraction. 98

4.4 A sample of the natural images from ImageNet 2012 used for training both OverFeat and VGG-S. 99

4.5 Area under the average ROC curves obtained from 200 trials [58] using OverFeat features with ℓ_1 or ℓ_2 regularized logistic regression, respectively. First row: using images without vessel inpainting. Second row: using images with vessel inpainting. The blue line corresponds to the best AUC value for the row. 103

5.1 Examples of red lesions observed in fundus photographs from DIARETDB1 [97]. 108
5.2 Overview of our method for red lesion detection. 111
5.3 Red lesion candidate detection. See Section 5.2.1 for a detailed description of the process. 112
5.4 Effect of the FOV expansion on the lesion candidates located closely to the border of the FOV. 112
5.5 CNN training set. Random sample of 200 patches for (a) non lesions and (b) true lesions. See Section 4.3.1 for details of the construction of the training set. 115
5.6 Feature based on vessel segmentation. (a) I. (b) I with candidates superimposed. (c) Vessel segmentation. (d) Vessel segmentation after removing spurious elements. (e) Vessel segmentation after morphological closing. 120
5.7 Per lesion evaluation in Experiment 1. FROC curve and CPM values obtained on the DIARETDB1 test set. 126
5.8 Per lesion evaluation for each lesion type in the DIARETDB1 test set. 127
5.9 Per lesion evaluation in Experiment 2. FROC curve and CPM values obtained on e-ophtha. 128
5.10 Qualitative results. (a) image015 from the DIARETDB1 test set. (b) Ground truth labeling at a $> 25\%$ level agreement. (c) Red lesion detections obtained by thresholding the probabilities at 0.644, which corresponds to an average FPI value of 1. (d) Detail from (c) showing lesions unlabeled on the ground truth but identified by our method. 130
5.11 Per image evaluation. ROC curves for (a) DR screening (R1 vs. R2, R3 and R4) and (b) need for referral (R1 and R2 vs. R3 and R4) on the MESSIDOR data set. 131

5.12 Per image evaluation on e-ophtha. ROC curve for DR screening. . . . 132
5.13 Learned filters on the first layer of our CNN, as obtained for each experiment in Table 5.4. 134
5.14 The t-SNE visualization of the patches from DIARETDB1 test set as mapped using the deep learned features, the hand crafted features and our hybrid feature vector. Left side: color coded labels for each test sample. Right side: patches around the candidates, as visualized using the t-SNE mappings. Details for different types of lesion candidates are shown in Figure 5.15. 135
5.15 Details from the t-SNE visualization in Figure 5.14 for different types of red lesion candidates (true lesions, vascular structures, speckles of dirt in the lens and false detections in vessel curves in the optic disc). 136

List of Tables

2.1 Classification of computer-assisted tools for screening and diagnostic of DR and glaucoma. 47

3.1 Evolution of F1-score during forward feature selection on DRIVE. . . 77
3.2 Quantitative evaluation of the results obtained on DRIVE, STARE, CHASEDB1 and HRF, using only the unary potentials (UP) or the fully connected CRF (FC-CRF). 79
3.3 Comparison of average Se, Sp, Pr, $F1$-score, G-mean and MCC values of our method with respect to other existing blood vessel segmentation algorithms and the 2nd human observer, when evaluating on DRIVE. 86
3.4 Comparison of average Se, Sp, Pr, $F1$-score, G-mean and MCC values of our method with respect to other existing blood vessel segmentation algorithms and the 2nd human observer, when evaluating on STARE. 87
3.5 Comparison of average Se, Sp, Pr, $F1$-score, G-mean and MCC values of our method with respect to other existing blood vessel segmentation algorithms and the 2nd human observer, when evaluating on the CHASEDB1 and HRF. 89

5.1 CNN architecture. Convolutional layers (conv) indicate width, height and depth of each learned filter. Pooling layers (pool) include the dimension of the pooling operation and the stride. Dropout is only applied after the first convolutional layer with a low dropout probability. 117
5.2 Summary of the hand crafted features used to complement our CNN. 119

5.3 Distribution of DR grades in the MESSIDOR data set, and diagnostic criterion. MA = microaneurysms, HE = hemorrhages and NV= neovascularizations. 121
5.4 Experimental setup. β is the value for the balanced cross-entropy loss (Equation (5.5)). 123
5.5 CPM values and per lesion sensitivities at FPI= 1 for Experiments 1 and 2. 129
5.6 Comparison of DR screening and need of referral performance on the MESSIDOR data set. Se values correspond to those obtained at a $Sp = 50\%$. 133

Chapter 1

Introduction

In this thesis we present novel methodological contributions to the field of automated fundus image analysis. In particular, the tools that we propose are based on machine learning techniques, and are applied for the diagnosis of two diseases that are among the major causes of avoidable blindness: diabetic retinopathy and glaucoma. In Section 1.1 we briefly describe the motivations of our work, while in Section 1.2 we provide a general explanation of machine learning and supervised learning techniques. Finally, Section 1.3 summarizes our contributions, the publications derived from our work and the thesis outline.

1.1 Motivation

According to the World Health Organization (WHO), the estimated number of people visually impaired in the world is 285 million, with 39 million blind and 246 million having low vision [149]. Moreover, 65% of people visually impaired and 82% of all blind are 50 years and older. With the natural aging of the global population, it is expected that the prevalence of these cases will increase in the next years [149]. In 2010, the WHO reported that more than 26 million people in Latin America suffered from visual impairment, and 3.2 million were blind [54]. However, 80% of these cases could have been prevented or treated through medical interventions if they would have been detected in time [54]. Similar characteristics were observed in Europe as well [164].

Several large-scale population-based studies and meta-analysis have concluded that diabetic retinopathy (DR) and glaucoma are among the three most common causes of visual impairment and/or blindness in the world, jointly with age-related macular degeneration [106, 194]. In Europe, for instance, these conditions explain approximately 30% of the blindness cases [164], while in Latin America the proportion of blindness due to macular degeneration, glaucoma and DR has significantly increased over the last years [107]. Compared to the global averages, Latin America showed more blindness due to age-related macular degeration, diabetic retinopathy and glaucoma than as a consequence of refractive errors [107].

Fundus photographs are a medical imaging modality that is extensively used for the detection of diabetic retinopathy and glaucoma. This is in part due to their non-invasive nature and the simplicity and low cost of their acquisition process [2]. However, the manual analysis of these images is prohibitive in the context of large scale screening campaigns, where numerous images must be analyzed in a relatively short time [3]. Furthermore, the diagnostics may vary significantly from one physician to another, depending on their professional experience, the quality of the images, the time devoted to the task or even fatigue. In an effort to mitigate these issues, methods for computer-aided diagnosis of ophthalmic diseases based on fundus images are being actively explored. In such a scenario, a computer takes an input image and apply a series of algorithms to provide feedback to physicians, including a labelling of the regions of interest or a probability indicating the level of risk of the patient from suffering the disease. This information allows to improve clinicians consistency and accuracy, also reducing the intra-expert variability [3].

As we will see in Chapter 2, there exists several strategies for automated diagnosis of DR and glaucoma. However, these models still suffers from many issues that must be solved in order to improve their performance on real, clinical scenarios. In the following we analyze the issues that the methodological contributions of this thesis aim to deal with:

- In general, automated systems for computer-aided diagnosis from fundus images rely on a first step that consists on segmenting the retinal vasculature [63]. As blood vessels are the most evident anatomical structure in fundus photographs, they are frequently used as a reference for detecting and/or seg-

menting other structures such as the optic disc–which is relevant for glaucoma assessment–or pathological red lesions–which are associated with DR–. Moreover, the architectural distribution of the retinal vasculature provides valuable information that is usually assessed by the physicians to determine if patients suffers or not from a certain ophthalmic disease. Current methods for automated blood vessel segmentation usually fail to deal with the narrower vessels, mainly due to their limitation to incorporate prior knowledge about the shape of the underlying structures. Consequently, automated systems are negatively affected by these failures. Hence, having better algorithms for retinal blood vessel segmentation is essential to improve the effectiveness of these screening systems [63].

- Most current methods for glaucoma detection are based on quantifying properties of the optic nerve head, which is the anatomical part of the retina mostly affected by the disease [75]. The main issue with this strategy, though, is that it requires the delineation of the optic disc and its internal parts. Furthermore, other valuable information such as the distribution of the retinal vessels is ignored by this approach. In contrast, deep learning methods are able to learn by themselves a series of features that can be associated to the disease, without requiring a manual engineering process [108]. However, this is achieved at the cost of using large amounts of annotated data during the training phase. Thus, developing novel methods for glaucoma assessment that are independent of the optic disc segmentation process is valuable, although taking into account that they should be able to exploit features from other anatomical or pathological parts of the retina.

- Most of the available systems for automated DR screening require the detection of red lesions, which are the earliest signs of the disease but are difficult to detect by human observers without an intense, time-consuming analysis of the images. Current baseline methods for automated red lesion detection are based on hand crafted features, which usually suffers from a limited ability to generalize the lesions characteristics properly. As we mentioned before, deep learning based strategies are a suitable alternative to avoid the manual designing of features, but their applicability is limited by the amount of training

data. The cost of collecting data to learn classifiers for red lesion detection is extremely high, as it requires to manually label these structures. Hence, it is important to develop new strategies for detecting red lesions using deep learning methods, but considering the fact that they should be able to learn how to identify the structures from data sets with a limited number of annotated lesions.

1.2 Machine learning

As previously stated, this thesis applies machine learning techniques for computer-assisted diagnosis of ophthalmic diseases. The main reason behind this decision is that machine learning provides a natural framework for developing algorithms for this task. Most of the currently available systems apply variants of these methods in many of their stages, including feature learning and extraction, image or lesion classification, segmentation and so on [2]. Although more details will be provided in the following chapters, this section briefly introduces the field and describes the general idea behind these methods.

Machine learning is a general term that groups a large variety of algorithms that are able to learn how to perform a given task without requiring to be explicitly programmed [77]. Although it is generally considered as a subfield of computer science, machine learning is built on top of other knowledge areas such as statistics, computational theory and optimization. The general idea behind these strategies is to train a model from a given set S of labelled or unlabelled samples $\mathbf{x}$, which were obtained from the domain of the problem to solve. By optimizing an objective function over the training set S, a set of parameters θ is learned. This model is afterwards applied in test time to classify new samples.

In this thesis we made use of supervised models for pixel or image classification [77]. Supervised learning is based on models learned from training sets S that are composed of pairs $(\mathbf{x}^{(i)}, y^{(i)})$, $i = \{1, 2, ...|S|\}$, where i is the number of training samples. Each pair is made up of a feature vector $\mathbf{x}^{(i)} \in \mathcal{X}$ that describes the sample by means of a series of quantitative measurements, and its corresponding ground truth label $y^{(i)} \in \mathcal{Y}$, with $\mathcal{X}$ and $\mathcal{Y}$ the sets of all possible features and labels that

might be associated to any sample. Depending on the application domain, the labels could indicate if a pixel belongs to a region of interest, for instance, or if an image corresponds to a healthy patient or to someone suffering from a disease.

A classification function $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ is learned from $S = \{(\mathbf{x}^{(1)}, y^{(1)}), ..., (\mathbf{x}^{(n)}, y^{(n)})\}$ by minimizing the empirical risk:

$$\mathcal{R}_S = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y^{(i)}, f_\theta(\mathbf{x}^{(i)})) \tag{1.1}$$

where θ is a set of parameters of the function f, and $\mathcal{L}$ is a loss function that compares the estimated output $\hat{y}^{(i)} = f_\theta(\mathbf{x}^{(i)})$ with respect to the ground truth label $y^{(i)}$, and is designed according to the problem.

The supervised learning framework can be used for image classification or segmentation by just adapting the loss function, the feature set and the input labels accordingly [77]. This setting allows us to apply these models for the tasks of blood vessel segmentation or for DR or glaucoma screening. Further details about how we applied these strategies in our contributions will be provided in the next chapters.

1.3 Contributions and thesis outline

This thesis begins with an extensive description of the current knowledge regarding automated detection of DR and glaucoma from fundus images. We first provide insights about the anatomy and physiology of the human eye, with special emphasis in the retina and their anatomical components. We then describe the diseases, underlying their earliest signs and how they are manually identified by physicians using different imaging modalities. We also provide a classification of the mainly used approaches for automated detection of both diseases, discussing their advantages and disadvantages.

Our first contribution is described in Chapter 3, and consists on a retinal blood vessel segmentation method based on learning fully connected conditional random field models using a structured output support vector machine. Pixel classification approaches based on supervised learning have been developed to solve this task, although they are not able to incorporate shape priors within the learning process.

To overcome this difficulty, we propose to use conditional random fields with high order pairwise potentials, which are able to better model the interaction between the pixels that comprise the vessels. We take advantage of recent advances on inference in fully connected conditional random fields [102], and integrate them on a structured output learning framework that has been extensively used in computer vision [151]. We validate this approach on benchmark data sets, resulting in better performance than other currently existing methods when evaluated in terms of global accuracy measures [154]. This chapter is based on the following publications:

- J. I. Orlando and M. del Fresno. Reviewing preprocessing and feature extraction techniques for retinal blood vessel segmentation in fundus images. *Mecánica Computacional*, XXXIII(42):2729–2743, 2014.
  This article presents an extensive analysis of different preprocessing and feature extraction techniques for retinal vessel segmentation in fundus images. The surveyed characterization approaches are evaluated in terms of their contribution to the improvement on the outputs of a support vector machine classifier.

- J. I. Orlando and M. B. Blaschko. Learning fully-connected CRFs for blood vessel segmentation in retinal images. In P. Golland, C. Barillot, J. Hornegger, and R. Howe, eds., *MICCAI 2014, LNCS*, volume 8149, pp. 634–641. Springer, 2014.
  This paper introduces for the first time our segmentation method based on fully connected conditional random fields. Results on the benchmark data set DRIVE are provided, and a comparison with respect of other state of the art approaches is presented as well, showing that our method achieves a competitive performance with respect to a second human observer.

- J. I. Orlando, E. Prokofyeva, and M. B. Blaschko. A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images. *IEEE Transactions on Biomedical Engineering*, 64(1):16–27, 2017.
  This paper extends our previous article with further experiments on other data sets such as STARE, CHASEDB1 and HRF. It also incorporates a scaling factor that is applied on feature parameters to adapt them to deal with higher

resolution images.

In Chapter 4 we introduce a novel strategy for glaucoma identification based on transferring convolutional neural networks (CNNs). The unavailability of large, annotated data sets of fundus images of patients with and without glaucoma makes unfeasible to train CNNs from scratch to solve this task. As an alternative, we propose to transfer CNNs that were pre-trained from non-medical data to glaucoma detection, under the hypothesis that CNNs trained from large enough databases of natural images are able to extract generic enough features. By first adapting fundus images using state of the art preprocessing techniques such as vessel inpainting, contrast enhancement and zooming on the optic nerve head, we extract CNN based features that are afterwards used to train a logistic regression classifier. Such an approach is able to achieve competitive performance with respect to other state of the art methods on benchmark data sets. This chapter is briefly based on the following paper:

- J. I. Orlando, E. Prokofyeva, M. del Fresno, and M. B. Blaschko. Convolutional neural network transfer for automated glaucoma identification. volume 10160, pp. 101600U–101600U–10. 2017. doi: 10.1117/12.2255740. URL `http://dx.doi.org/10.1117/12.2255740`.

Our third contribution is a red lesion detection strategy for DR screening, which is extensively described in Chapter 5. It consists on first retrieving a large number of lesion candidates with an unsupervised method, based on morphological operations. These structures are afterwards classified into true and false positive lesions by means of a hybrid feature vector of hand crafted and deep learned descriptors. A light CNN architecture is trained from scratch using patches around lesion candidates and their corresponding per lesion annotations. The features learned by the network are then augmented using other hand crafted descriptors. This ensemble strategy is used to train a Random Forest classifier, which provides a per-candidate probability that indicates if it corresponds to a true red lesion or not. Results obtained on benchmark data sets shows that this strategy allows to improve the original performance of the CNN, also surpassing other state of the art approaches both on a per-lesion and a per-image evaluation. The chapter is based on the following article:

- J. I. Orlando, E. Prokofyeva, M. del Fresno, and M. B. Blaschko. Learning to detect red lesions in fundus photographs: An ensemble approach based on deep learning. *arXiv preprint arXiv:1706.03008*, 2017.

Finally, we conclude this thesis in Chapter 6, summarizing our main results and achievements, and providing further research lines that can be derived from our original contributions.

Chapter 2

Background

This chapter summarizes the essential aspects of automated fundus image analysis for the assessment of retinal diseases. Section 2.1 describes the eye's anatomy, with special attention to the retina and its function. In Section 2.2 we analyze the two diseases in which this thesis is focused on, diabetic retinopathy and glaucoma. Section 2.3 details the main imaging modalities for detecting these diseases. Finally, Section 2.4 presents a brief review of the current state of the art on computer-assisted diagnosis of retinal diseases.

2.1 The eye and the retina

The vision sense is one of the most important senses in humans, and is achieved by a complex system that involves mainly two organs, the eyes and the brain. The eyes are responsible of focusing the light of the environment and convert it into electrical impulses that are afterwards interpreted by the brain [177].

The eye (Figure 2.1) is a spherical structure that is divided into two main parts, the anterior and posterior segments.

The anterior segment is visible from outside, and accounts for one-sixth of the entire eye. It is composed by the cornea, the pupil, the iris and the lens [96]. The cornea is a transparent tissue located at the outermost part of the anterior segment. The pupil is an opening that is seen as a black circle surrounded by a colored region, the iris, which is made of smooth muscle tissue that can contract and dilate the

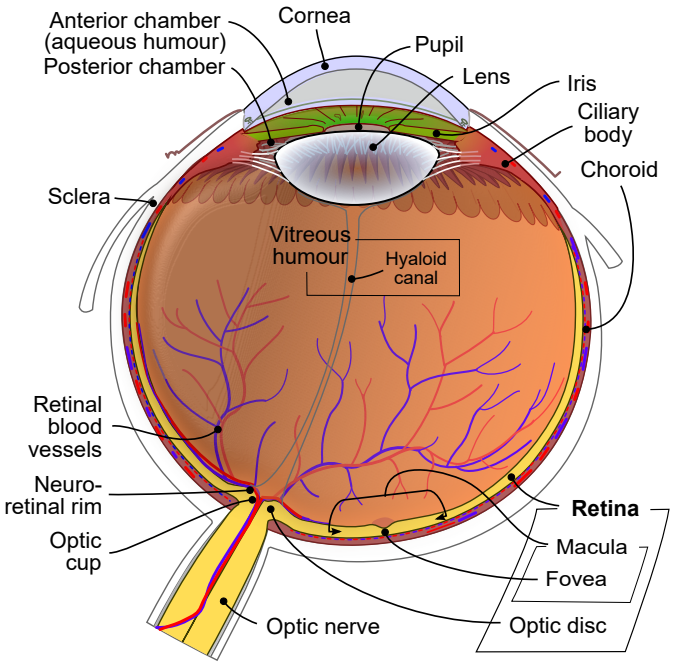


Figure 2.1: Schematic representation of the eye. Source: [42].

pupil. This part automatically controls how much light enters into the eye, allowing less when the environment is sufficiently illuminated or more when it is darker [96]. The lens is a convex, transparent disc that helps to focus the light. Between the iris and the lens, and the cornea and the iris, there is a transparent fluid known as aqueous humour. It is secreted from the ciliary epithelium, a structure supporting the less, and it outflows through a drainage system whose main part is the trabecular meshwork. Its function is to maintain the intraocular pressure and inflate the eyeball to sustain its almost spherical shape, but it is also responsible of providing nutrition to avascular tissues such as the posterior cornea, the trabecular meshwork, the lens and the anterior vitreous [96].

The posterior chamber involves the remaining five-sixth of the eye, and is surrounded by a layer of protective fat in which a set of extrinsic eye muscles are connected [96]. The posterior chamber is composed of the vitreous and a wall made up of three different layers of tissue, a fibrous, a vascular and an inner layer. The vitreous is a jelly like substance that fills the entire posterior cavity and is respon-

sible of ensuring the spherical shape of the eye. The fibrous tunic surrounds both the anterior and posterior segments, and is made mostly of connective tissue. It is composed by the cornea and the sclera, which is a shell of white colored material. The vascular layer corresponds to the choroid, and is composed of a network of thin capillaries that supply blood and oxygen to the other layers [96]. Finally, the inner layer is the retina.

The retina is a thin tissue that is partially transparent and is composed of millions of photoreceptors cells and neurons [177]. There are two types of photoreceptor cells, namely rods and cones. Cones are able to distinguish fine details and colors, and are adapted to be exposed to more light that the rods. Rods, on the contrary, only register grayscales and are more relevant for peripheral vision and for sight under low light conditions. Both types of cells are connected with the ganglion neurons by a layer of bipolar cells. Ganglion neurons transmit the electrical signals through the optic nerve to finally reach the brain [203].

Anatomically, the retina is briefly composed of three structures: the macula, the retinal vessels and the optic disc [2]. The macula is a yellow region which is specialized on the fine detail view. It has an approximate size of 5 mm of diameter, and is located in the posterior part of the retina. Histologically, the macula is characterized by the total absence of rods and the existence of a large amount of cones. At the center of the macula there is a small depression known as the fovea, which is darker in color and has an approximate extension of 1.5 mm of diameter [177]. The optic disc–also known as the optic nerve head, ONH–is a circular area located at the center of the retina, where the optic nerve enters into it. This area is about 1.5×2.5 mm approximately in adults, and is a blind spot as it does not have any cones and rods. Inside the ONH there is a small physiological excavation known as the optic cup (OC), which is observed as a pale area inside the neuroretinal rim [177]. Around 65% of the blood needed by the retina is supplied by the choroid. However, the remaining 35% is given by the retinal vessels. These thin arteries and veins visually appear as elongated features spanning from the center of the optic disc.

Vision is achieved by means of the interaction of all these anatomical parts [96]. First, the light reflected from objects enters into the eye through the cornea, which

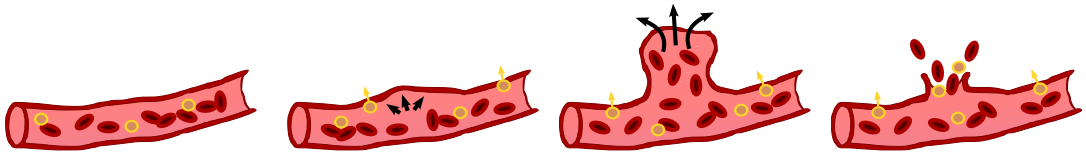
focuses the light to allow sharper and clearer vision. Depending on the illumination of the environment, the pupil will be dilated or contracted by the iris. After passing through the pupil, the light traverses the lens, which is flattened or expanded to better distinguish far or close objects, respectively. Finally, photons are projected into the retina, where the cones and rods convert the electromagnetic waves from the light into the electrical signals that are transmitted through the ganglion cells to the optic nerve, to be further processed by the brain [203].

2.2 Retinal diseases

Several ophthalmological and systemic diseases manifest themselves through changes on the retina [2]. Moreover, some of them can affect the vision itself if left untreated [2]. In this thesis we will focus on two specific conditions, namely diabetic retinopathy and glaucoma.

2.2.1 Diabetic retinopathy

Diabetic retinopathy (DR) is one of the most common consequences of diabetes. Diabetes mellitus is a systemic disease in which glucose cannot be properly metabolized due to issues in the insulin generation or usage [131]. Two different types of diabetes can occur, Type 1 and Type 2. Type 1 diabetes is characterized by the immune system mistakenly attacking and killing cells of the pancreas. As a consequence, only small or null amounts of insulin are released into the body. This condition accounts 5 to 10% of the diabetic cases, and is usually developed in childhood or adolescence. Type 2, on the contrary, is caused by a syndrome known as insulin insensitivity, in which the body cannot properly use the insulin or do not produce enough. Most of the patients with diabetes suffers from this condition. The main consequence of both types of diabetes, as outlined above, is an improper glucose metabolism. This might lead to hyperglycemia, a condition in which an excessive amount of glucose circulates in the blood plasma. Hyperglycemia can lead by itself to a wide variety of vascular complications if left untreated for long periods of time [2]. Recent studies have shown that the number of people suffering from diabetes is expected to increase from 366 million (2.8%) in 2011 to 552 million (4.4%) by 2030 [215]. Moreover, the



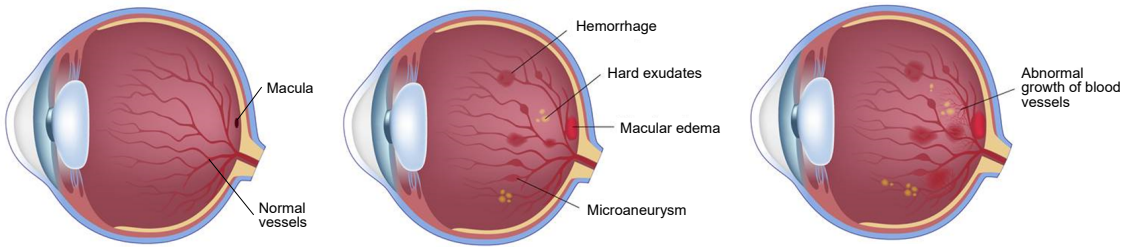
(a) (b) (c) (d)

Figure 2.2: Microvascular complications due to hyperglycemia. (a) Healthy microvessel. (b) Hyperglycemia weakens vessel walls and increments the blood flow, while lipids might be leaked. (c) Weaker vessel walls and the increase shear strees favours the formation of microaneurysms. (d) The increased pressure can eventually lead to the rupture of the microaneurysm and produce hemorrhages.

greatest increase is expected to occur in low and middle-income countries [215]. In Argentina, prevalence of diabetes has grown from 8.4% to 9.6% from 2005 to 2009 [175], and is among the main causes of death [6]. Although the causes of diabetes are not entirely explained yet, it was observed that obesity, a sedentary lifestyle or a genetic background are all associated with an increased risk of developing the disease [131].

DR is one of the most common consequences of microvascular damage due to diabetes [2], and is one of the leading causes of preventable blindness in the world [164]. About 5% of people with Type-2 diabetes have DR, and it is expected that the number of patients suffering this disease will significantly increase in the next years [3]. In Argentina, recent studies identified DR as the second cause of avoidable blindness, accounting for 16% of the cases [23].

Figure 2.2 describes the microvascular complications associated to DR. Hyperglycemia damages retinal pericytes, which are specialized cells located in the surface of capillary vessels that control the microvascular blood flow [195]. As a consequence, the vessel walls are weaken and became leaky, and the blood flow is increased (Figure 2.2.1). This last factor increments the shear stress in the vessel walls [120], favouring the formation of microaneurysms (MAs) [131] (Figure 2.2.1). MAs are balloon-shaped deformations on the vessel walls of the retinal microvessels that can break and produce leakages of blood on the retinal layers, known as hemorrhages (HEs) (Figure 2.2.1). The most commonly used term to refer to both MAs and small HEs is "'red lesions"', as they are seen as small red dots in fundus images [143, 180].



(a) No DR (b) Nonproliferative DR (c) Proliferative DR

Figure 2.3: Diabetic retinopathy stages. Source: [19].

On the other hand, the permeability of the vasculature ends up allowing the proteins and lipids in the blood to escape from vessels. These accumulations are known as hard exudates, and, jointly with red lesions, are the main clinical findings associated with nonproliferative DR, the earliest stage of the disease [2].

The microvascular complications described above have consequences that eventually affects the retinal metabolism (Figure 2.3). When the number of red lesions augments, larger regions of the retina stops receiving enough blood supply, causing a shortage of the oxygen and glucose needed to keep the tissue alive [229]. This phenomenon is known as ischemia, and the body reacts to it by releasing an angiogenic factor that stimulates the generation of new vessels to bypass the damaged ones. This neovascularization process conduces to the most advanced stage of DR, known as proliferative DR (PDR) [214]. At this point, the new weak vessels can also break and produce new leakages, increasing the risk of irreversible vision loss (from 25.6% to 36.9%) due to retinal detachment and vitreous hemorrhage [2].

Medical treatments depends on the progression of the disease. In general, diabetic patients with no evident DR are given drugs to control the amount of glucose in blood [131]. Additionally, they are required to attend to the ophthalmologist at least once an year for an eye exam, to look for the existence of red lesions. In the more advanced stages, however, the treatment is limited [40]. If severe neovascularization is observed, laser photocoagulation and vitrectomy can allow to prevent visual loss [40]. Nevertheless, both treatments have an associated risk of additional vision loss and cannot revert visual acuity loss [40]. In other cases, patients receive intravitreal injections to avoid vascular proliferation [40].

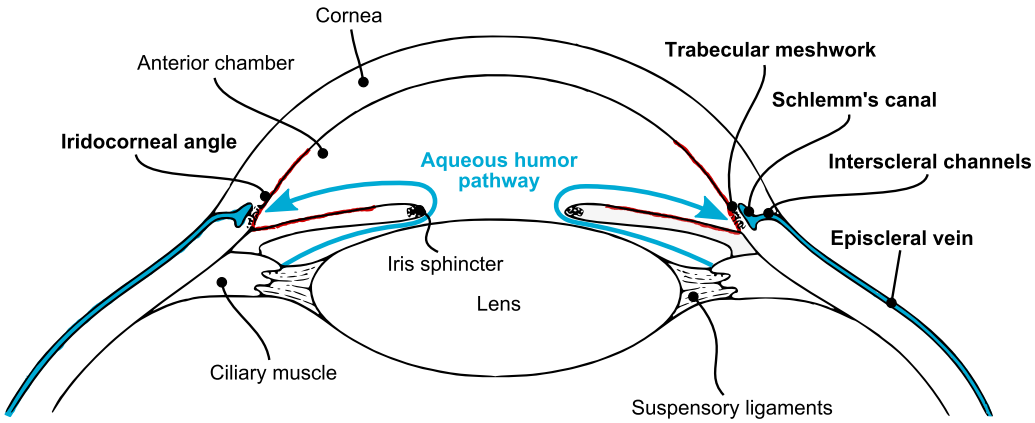


Figure 2.4: Aqueous humor flow. In light blue, aqueous humor pathway. In red, iridocorneal angle.

Early detection of DR is essential to prevent its progression and to reduce sight threatening. However, DR is asymptomatic in its first stages, so the only way to detect it is by looking for the clinical signs associated with the disease [2]. The earliest sign of DR are red lesions, which are assessed by physicians by means of an ophthalmoscopy or an imaging modality such as fundus photographs (Section 2.3) [2, 158]. Insights about DR screening based on fundus images are given in Section 2.4.

2.2.2 Glaucoma

Glaucoma is a chronic and irreversible neuro-degenerative condition that is one of the leading causes of preventable blindness in the world [194]. In 2013, the number of people aged 40-80 years with glaucoma worldwide was estimated to be 64.3 million, and it is expected to increase to 76 million in 2020 and 111.8 million in 2040 [194]. In European countries, glaucoma explains 20.5% of the cases of avoidable blindness [164]. Moreover, the prevalence of blindness due to glaucoma continued to increase in Latin America over the last decade [107].

Although the term glaucoma refers to a family of different diseases, all its variants are characterized by the damage of the ONH due to high pressure in the eye. This is a consequence of an accumulation of aqueous humor due to a defect of its drainage system (Figure 2.4). The aqueous humor is secreted by the cilliary body and flows through the pupil to the anterior chamber, from where it drains through

the trabecular meshwork. This spongy tissue is located in the cornea, close to the iridocorneal angle–which is the angle between the cornea and the iris–, and its responsible of draining most of the aqueous humor [41]: the liquid is collected by the trabecular meshwork into the Schlemm's canal, from which it flow throw the interscleral channels to the episcleral vein [41]. In healthy patients, the secretion rate of aqueous humor is almost equivalent to the draining rate, so the pressure in the eye remains constant. However, if the secretion rate is higher than the draining rate, the aqueous humor starts to accumulate in the anterior chamber. This region will be ultimately filled by the liquid, and the posterior chamber will begin to saturate too, progressively elevating the intra-ocular pressure (IOP) and compressing the vitreous to the retina. This condition can potentially damage the nerve fiber layer, the vasculature and the ONH, leading to a progressive loss of vision than can ultimately result in irreversible blindness if left untreated.

Glaucoma risk factors include age, genetic background, high IOP, obesity, consumption of corticosteroids and eye traumas [111]. Moreover, people with a family history of glaucoma are recommended to assist to the ophthalmologist once an year to control the IOP and to assess if the ONH has suffered any damage [164].

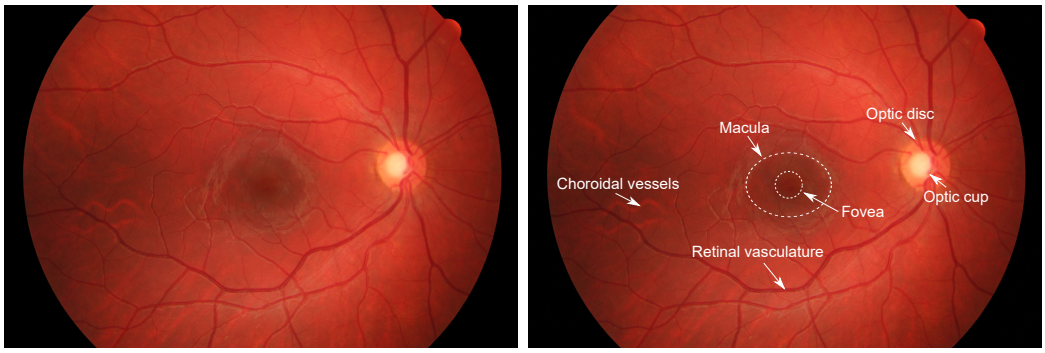
In general, glaucoma can be classified into two main diseases, namely open-angle and closed-angle glaucoma. Primary open-angle glaucoma (POAG) accounts for the vast majority of glaucoma cases, and occurs when the aqueous humor cannot be sufficiently drained due to the degeneration and obstruction of the trabecular meshwork [132]. This produces a chronic, painless growing of the IOP that is generally asymptomatic but can eventually lead to ONH damage and blindness. Closed-angle glaucoma is a less common disease in which the iridocorneal angle is suddenly blocked, and the aqueous humor cannot flow outside the anterior and posterior chambers. This condition is abrupt and lead to a painful increase of the IOP that can lead to permanent vision loss relatively fast [213].

POAG is known as the silent thief of sight, and is the most difficult type of glaucoma to detect [164]. This is due to the fact that patients suffering from the disease do not notice its effects until vision is reduced. In general, the first manifestations of vision loss due to glaucoma are related with blurred peripheral vision. As the disease progresses, however, the visual field is more decreased and patients end up

having a tubular vision that is eventually lost, too [213]. From now on we will use the term glaucoma to refer only to POAG, which is one of the diseases analyzed in this thesis.

At least half of patients with glaucoma remain undiagnosed, while more than half of those who are undergoing treatment do not have the disease [164]. As glaucoma is a chronic life long condition, the major challenge is to be able to screen for glaucoma to detect the large number of undiagnosed people [164]. Current large scale screening practice includes the examination of the ONH through fundus imaging and the assessment of retinal nerve fiber layer (RNFL) defects and the measurement of the cup to disc ratio (CDR) [192]. Despite the fact that RNFL defects are considered the earliest sign of glaucoma, the CDR is still the most commonly used diagnostic criteria for screening glaucoma using fundus images [75]. In general, it is computed as the ratio with respect to area, vertical length and horizontal length of both the OD and the optic cup. When the pressure within the eye is too high, the optic nerve fibers are damaged and start to disappear. As a consequence, the optic cup becomes larger with respect to the OD, and causes an increase of the CDR value [75]. The manual measurement of this parameter is extensively used as a biomarker of glaucoma, although it is known to be influenced by factors such as sex, age or race, etc. [81]. Patients showing a high CDR value usually undergoes to other studies to complement the original one by assessing the nerve fiber layer or the retinal vessels using such imaging modalities as optical coherence or Heidelberg Retinal tomographies [216].

Currently there is no cure for glaucoma, although there exist different treatments to reduce its effects or to prevent its progress. Those treatments include medication, laser therapy and surgeries. Medications have the purpose of reducing the IOP, and are administered directly into the eye as drops. Laser treatments are used to deal with POAG, and consist in laser beams that are applied on the trabecular meshwork to stimulate the outflow of the aqueous humor. Surgeries are usually applied on the trabecular meshwork for opening new conducts to facilitate the drainage [213].



(a) Fundus image (b) Anatomical parts

Figure 2.5: Fundus image of a healthy subject, extracted from the HRF data set [145].

2.3 Retinal imaging

The anatomical characteristics of the retina allow physicians to examine it non-invasively using different imaging techniques [158]. Perhaps the most common way to perform this task is by means of an ophthalmoscopy, which consists on the observation of the fundus by means of an ophthalmoscope, an economical device consisting of a lens and a light source [222]. Retinal imaging is a robust alternative to ophthalmoscopy as it allows to preserve the examination as a digital image. The state of the art imaging modalities to analyze the retina are fundus photographs, fluorescence angiographies and optical coherence tomographies (OCTs) [2]. Despite the fact that this thesis is focused on using fundus photographs for computer-assisted diagnosis of retinal diseases, the other two imaging modalities are briefly described in the following sections.

2.3.1 Fundus photographs

Fundus images (Section 2.5), also known as retinographies or fundus photographs, are projective color images of the inner surface of the eye [157]. This imaging modality is the most economical technique for diagnosing retinal and systemic diseases, including DR and glaucoma [2]. The acquisition process is also non-invasive and relatively easy to perform [131]. Images are captured by means of a fundus camera (Figure 2.6), which consists mainly of two parts: a specialized low power microscope with a light

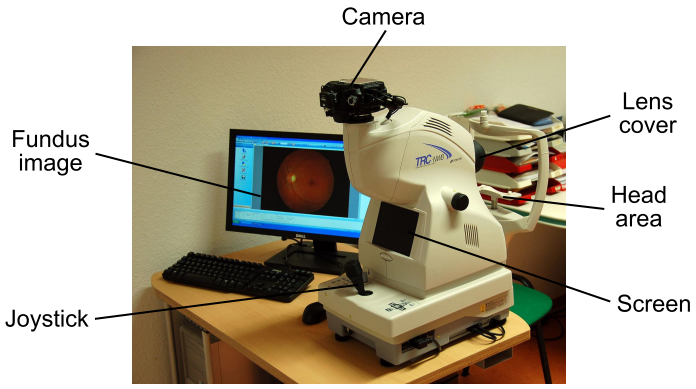
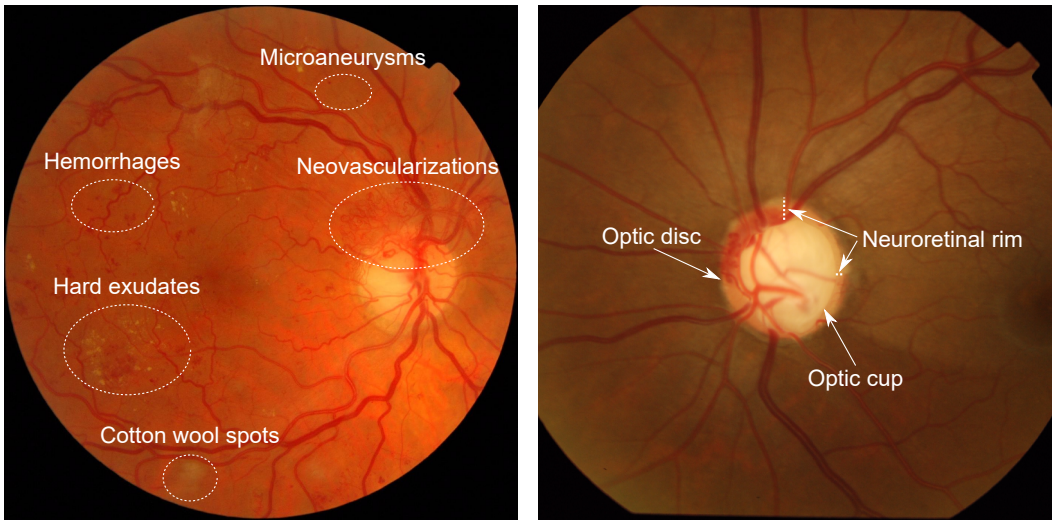


Figure 2.6: Fundus camera. Source: Ralf Roletschek / Wikipedia [171].

Figure 2.7: Volk iNview, a smartphone-based device for fundus image acquisition. Source: [148].

source that illuminates the interior part of the eye, and a conventional CCD digital camera that captures the photograph and stores it as a digital file. In the last years, however, portable devices such as those based on smartphones (Figure 2.7) have been introduced to reduce the burden of acquiring a fundus camera and to improve its mobility [47].

The capturing protocol using a conventional fundus camera starts with the patient resting her chin on a dedicated part of the device, and the operator aligning the camera with the iris and the pupil using a joystick. Then, the camera is zoomed to illuminate the posterior chamber, and a global vision of the retina is given on a screen. The patient then fixes her eyesight on a fixation target (usually a green light) and the operator manually adjust the focus to get a clearer vision of the fundus. Fi-



(a) Diabetic retinopathy. Source: MESSIDOR data set [49].

(b) Glaucoma. Source: DRISHTI-GS1 data set [187].

Figure 2.8: Fundus images for retinal disease diagnosis.

nally, a series of photographs of different parts of the retina are acquired by simply shooting the camera and asking the patient to fix the eye on different targets [2].

Figure 2.5(b) depicts the anatomical components of the retina as observed through fundus photographs. The retinal vasculature, the ONH, the macula and the fovea can be easily assessed using this imaging modality. Moreover, Figures 2.8(a) and 2.8(b) present fundus images of patients with DR and glaucoma, respectively. Pathological features such as red lesions, hard exudates, cotton wool spots or neovascularizations can be clearly observed in the DR case, while an increased size of the optic cup relative to the optic disc diameter is also evident in the glaucomatous patient. Compared with an ophthalmoscopy, fundus images allow a larger inspection of the retina [158]. Moreover, photographs of different parts of the retina can be combined using registration algorithms to cover a larger area than the originally given by the FOV of the camera [2].

The main disadvantage of fundus photographs is that they are 2D projective representations of the retina rather than 3D scans, making difficult to assess lesions occurring within the inner retinal layers or to measure the level of damage of the optic nerve. However, its low cost make them extremely useful to assess large

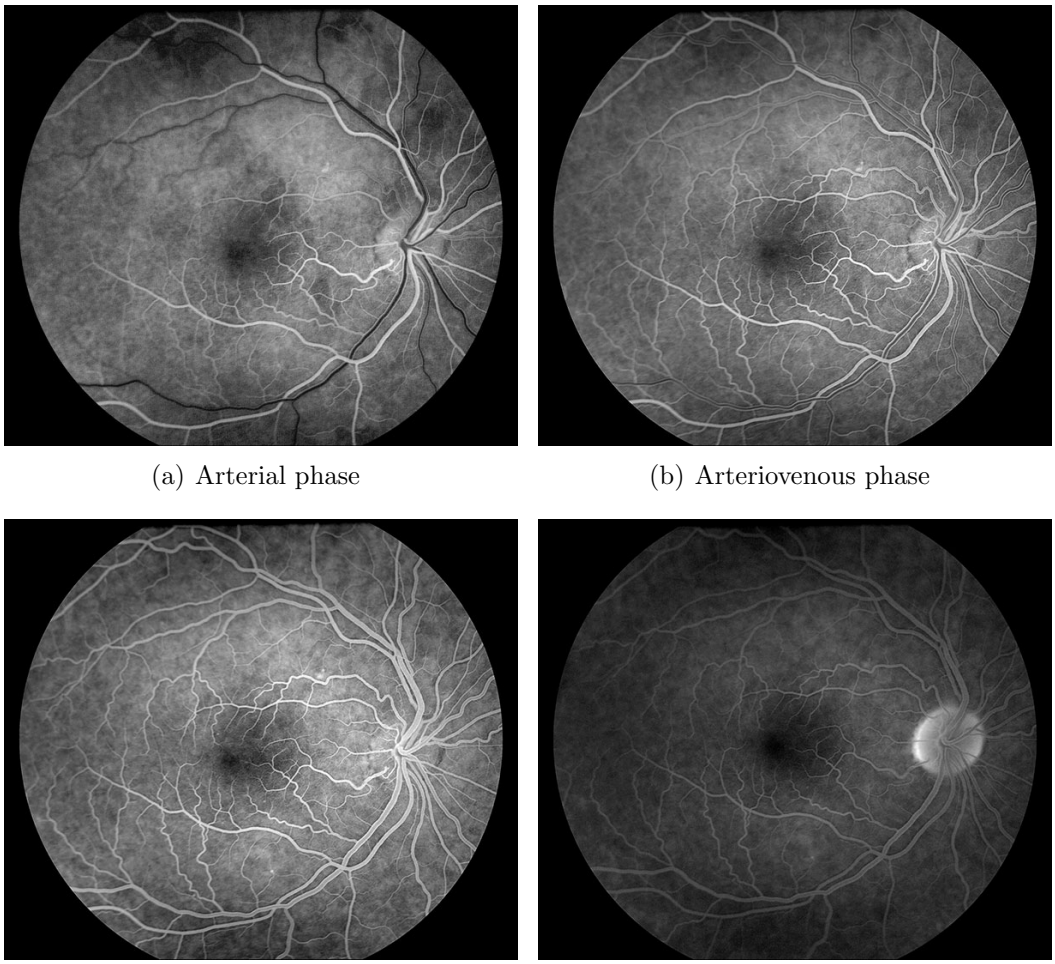
populations [2, 3].

2.3.2 Fluorescein angiography

Fluorescein angiographies (FA), also known as fluorescent or fundus fluorescein angiographies, allow to study retinal blood vessels not only from an anatomical point of view but also in terms of their functional behavior [2]. The fluorescence is a molecular property that allows certain compounds such as the sodium fluorescein to emit light a long wavelength when they are stimulated by a lower wavelength light [170]. By means of a fluorescent dye and a specialized angiographic camera, both the choroid and the retinal vessels can be assessed on different time intervals, providing photographs that describe the blood circulation on this layer [170].

The acquisition protocol is similar to the one used for fundus photographs. First, mydriatic drops are applied to dilate the pupil of the patient. Subsequently, an injection of sodium fluorescein is administered intravenously to the patient, and the angiographic camera takes a series of photographs that register the movement of the fluorescent dye through the vasculature [170]. This is achieved by capturing the light emitted by the compound when it is illuminated with a blue light. In general, four phases are photographed, namely the arterial, the arteriovenous, the venous and the elimination phases (Figure 2.9) [170]. During the arterial phase, the veins appear darker than arteries. In the arteriovenous phase both arteries and veins are distinguished as bright structures. In the venous phase, the arteries start to lose some brightness but veins still remain brilliant. Finally, the elimination phase comes between 5 and 10 minutes before injecting the dye, and is characterized by a lower contrast of the vessels with respect to the other retinal structures.

FA are mostly used to study DR patients and to assess macular edema or age-related macular degeneration [170]. MAs and exudates are more easily observed in FA than in fundus photographs, as the fluorescent dye is also leaked on this structures. This setting makes FA useful to aid laser treatments in which the sources of blood or lipids perfusion must be properly identified before. However, the injection of the fluorescent dye makes this procedure invasive, so other imaging modalities such as OCTs or fundus photographs are used, instead [2]. Patients can experiment nausea or vomits after the procedure, and cannot be screened using this imaging modality



(a) Arterial phase (b) Arteriovenous phase

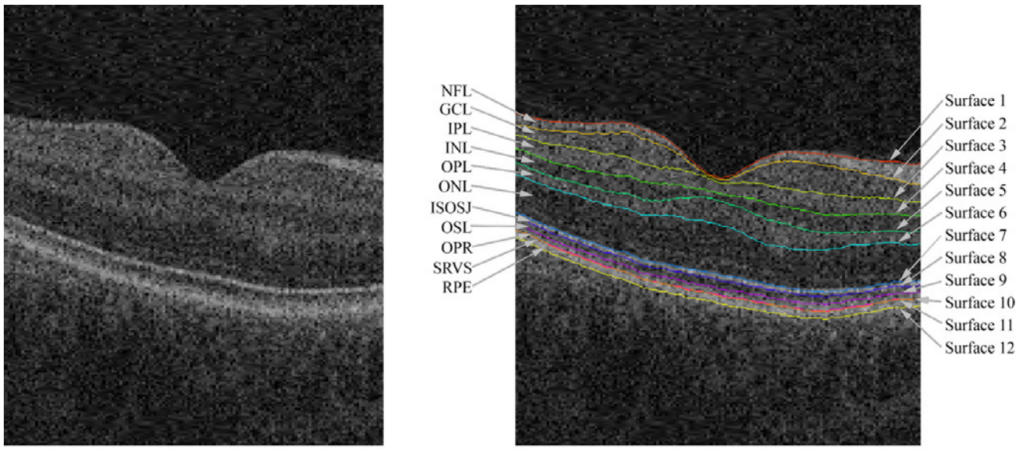
(c) Venous phase (d) Elimination phase

Figure 2.9: Fluorescein angiography. Phases. Source: [167].

if they have anaphylaxis or allergy to fluorescein or poor kidney function [170].

2.3.3 Optical coherence tomography

Optical coherence tomography (OCT) is a non-invasive, non-contact imaging technique that provides volumetric imagery of the subsurface tissues of the retina [191]. This is achieved by a principle known as low-coherence interferometry, which is based on a near-infrared light source that penetrates the retinal tissues and allows to recover cross-sectional scans with a resolution up to 10 times higher than the obtained



(a) (b)

Figure 2.10: OCT scan and retinal layers as obtained applying and automated segmentation method [112]. (a) X-Z image of the OCT volume. (b) Segmentation results, nerve fiber layer (NFL), ganglion cell layer (GCL), inner plexiform layer (IPL), inner nuclear layer (INL), outer plexiform layer (OPL), outer nuclear layer (ONL), inner segment outer segment junction (ISOSJ), outer segment photoreceptors (OPR), subretinal virtual space (SRVS-zero thikness in normals), and retinal pigment epithelium (RPE). Figure extracted from [191].

using ultrasound devices [2]. By combining each of this 2D slices it is possible to reconstruct a 3D volume of the retina that is suitable to assess different abnormalities in the vitreo-retinal interface, for instance, or deeper pathologies such as choroidal neovascularization, ONH damage due to glaucoma or retinal detachment [191]. It also allows to assess the retinal thickness, which is valuable for planning surgical or pharmacological interventions. Moreover, the OCTs scans provide not only an overview of the surface of the retina but also of its inner layers, including the nerve fibers, the pigment epithelium and the choroid [2].

Figure 2.10 presents an OCT scan of the macular region of the retina. The acquisition procedure is relatively simple, although it takes longer than a fundus photograph. Mydriatic drops are administered to the patients first in order to dilate the pupil, and the tomograph scans the retina during 10 to 14 minutes using an infrared beam. The time delay and magnitude change of such a low coherence light as it is backscattered by the tissues is analyzed by the device and transformed into gray

level intensities [59]. Several artifacts such as poor ocular media, patient compliance or saccadic movements can influence the manual assessment of these images by the experts [189]. Moreover, the speckle noise typical of this imaging modality makes its automated processing significantly difficult, as well [140].

2.4 Fundus image analysis for computer-assisted diagnosis

In this section we will summarize the most significant methods used in the literature for computer-assisted diagnosis of DR and glaucoma using fundus images. Detecting these two diseases at a stage when only a few manifestations are observed is a key task to perform, as early access to medical therapies is the only way to prevent these conditions to progress to blindness [127].

2.4.1 Overview

Fundus photographs represent the most economical imaging modality for DR and glaucoma screening due to several reasons. First, these images are obtained by means of a relatively simple acquisition process that can be performed by technicians rather than ophthalmologists. This means that little expert support is required to capture the images. This favors large scale screening campaigns based on telemedicine systems, in which expert physicians do not need to move to where the campaign is taken place [114]. Instead, technicians capture the images and transmit them through a secure network to a remote facility in which the physicians perform the diagnosis. Such systems are known to be cost effective and have been used for years to screen risk populations for glaucoma and DR [3, 136]. Smartphone-based fundus cameras have been integrated also to these schemes to improve both mobility and usability of the capturing devices.

The bottleneck on any screening campaign, however, is on the expert analysis of the images [3]. If the population to screen is significantly big, a large amount of fundus photographs will be retrieved and several ophthalmologists will be needed to diagnose each of the images. This is a tedious and time consuming task that can be

influenced by several factors such as the specialists fatigue or expertise, incrementing the inter- and intra-expert variability [3].

Computer-assisted methods for detecting these diseases are known to reduce the effect of these issues and the overall burden of the screening campaigns [127]. By means of systems that detect lesions or pathological changes, an ordering of the most critical cases can be automatically made so physicians decide which of them to analyze first [127]. Moreover, this information aids the experts to assess subtle changes that are difficult to observe with the naked eye such as those in the overall vascular distribution or red lesions [127].

In general, automated methods for screening ophthalmic diseases can be organized in four different categories, as indicated in Table 2.1. Vessel characterization approaches correlate changes in the architectural distribution of the retinal vasculature with the existence of the disease. Methods based on detecting pathological structures identify the diseases by quantifying lesions that are associated with their existence. Other strategies extract global image features to characterize both healthy and pathological scans to discriminate them. Finally, deep learning based models use convolutional neural networks to automatically learn the characteristics of the diseases. This classification is not strict, as several existing methods might overlap with more that just a single category. For example, the method for glaucoma screening by Chakravarty *et al.* [32] combines both image and segmentation based features to learn the classifier. Nevertheless, this classification groups most of the techniques currently available in the literature. The general ideas behind each of the approaches are given in the following sections.

2.4.2 Vessel characterization approaches

Several studies were conducted in the last decades trying to correlate changes in the architectural distribution of retinal blood vessels with DR or glaucoma. These studies are usually based on complex methods for vessel segmentation and characterization, or use commercial tools such as SIVA (Singapore I Vessel Assessment, [37], Figure 2.11) or VAMPIRE (Vessel Assessment and Measurement Platform for Images of the REtina, [159]). These systems first delineate the retinal vessels using an automated approach, and then extract a series of features that describe their

Table 2.1: Classification of computer-assisted tools for screening and diagnostic of DR and glaucoma.

	Vessel properties	**Detection of pathological structures**	**Global image features**	**Deep learning methods**
DR	Caliber [44, 137] Fractal dimension [38, 72]	Red lesions (MAs and HEs) [168, 176, 179], bright lesions (exudates) [176, 179], neovascularizations [110, 174]	Multiscale AM-FM [7], HOS [4], Texture (GLCMs and RLMs) [5]	CNNs [163], CNNs with abnormalities visualization [69, 165, 212]
Glaucoma	Caliber [11, 13, 90, 129], tortuosity [218], branching angle [218].	CDR [75, 91], RNFL defects [133, 146], PPA [117]	Raw intensities [26], texture [26, 130], FFT coefficients [26, 27], histogram [26], B-spline coefficients [27], wavelets [50].	CNNs [34] (limited by the no availability of public data sets).

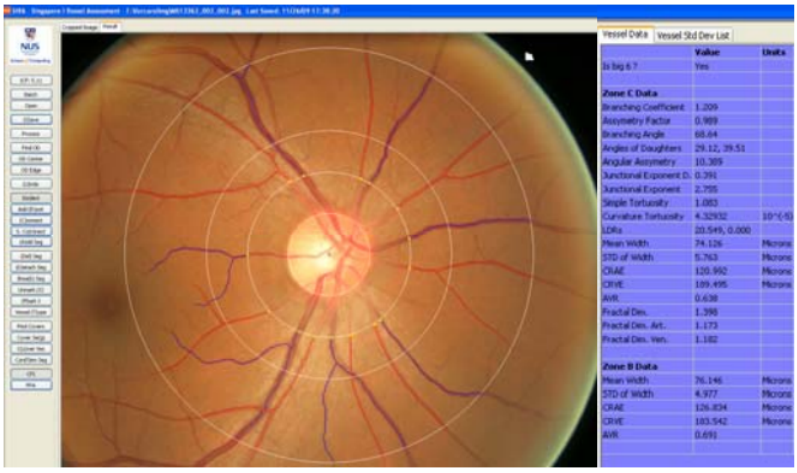


Figure 2.11: Screenshot of the Singapore I Vessel Assessment (SIVA) tool. Source: [1].

morphology and distribution with respect to other anatomical structures such as the macula or the ONH. In some cases, these features are only obtained from a small region located at a certain distance of the ONH, or from all the segmented vasculature. Moreover, some approaches are based on characterizing arteries and veins separately. In that case, classifying the vessels after segmentation is required.

The retinal vasculature can be characterized in terms of several properties, including the diameter of arteries and veins, the amount and angles of bifurcations, the tortuosity and area of the segments, etc. [9, 122]. The overall architecture and

distribution of the vasculature can also be described in terms of its fractal dimension [37]. Fractals objects are structures that are characterized by self-similarity, which means that a pattern persists at multiple scales [55]. Branching structures such as the retinal vessels are natural examples of these objects [84].

In general, the ability of these methods is limited by the performance of the vessel segmentation approach that is used to delineate the vasculature [86]. Vessel segmentation is known to be a significantly difficult task that is still an open problem in the literature [63]. Moreover, results obtained using vascular features for screening ophthalmic diseases are significantly surpassed by other methods based on more robust descriptors. However, this research has a remarkable impact in the understanding of the diseases and the development of novel therapies or treatments [63].

2.4.2.1 Diabetic retinopathy

Vessel calibers and fractal dimensions were mostly identified as potential features to identify early DR or to predict progression to PDR. In [137], a larger retinal venular caliber was observed in diabetic patients with retinopathy compared with patients with no evident retinopathy. Greater retinal fractal dimension showed in [38] to be independently associated with early DR signs in Type I diabetes. In the study presented in [72], fractal features were measured from retinal images of diabetic patients to estimate their association with micro and macrovascular complications such as DR, neuropathy and nephropathy. Patients with lower fractal dimension were seen to be more likely to have PDR and neuropathy. Vessel calibre was observed as an important risk factor to predict the progression to PDR in [44], and similar results for fractals were also observed in the same study.

2.4.2.2 Glaucoma

Vessel properties have also been analyzed as potentials features for characterizing glaucoma, although with contradictory results. In general, only three vascular characteristics were found to be associated with glaucoma: the caliber, the branching angle and the tortuosity of the vessels. Patients with glaucoma have been reported to have significantly smaller arteriolar and venular calibers than those without the disease in several studies [11, 13, 90, 129]. However, prospective data from the

Rotterdam Study did not show an association between retinal vascular caliber and glaucoma [85]. Another study but on FA arrived to a similar conclusion [17]. Finally, decreased arteriolar and venular tortuosity, and narrower retinal venular branching angle were associated with glaucoma in [218].

2.4.3 Detection of pathological structures

Most of the approaches for computer-assisted diagnosis of DR and glaucoma are based on detecting pathological structures associated with the early stages of the diseases [3, 51]. In the daily clinical practice, physicians look for any signs of MAs or HEs to determine if the patients has or not DR [180]. Similarly, the damage of the RNFL or a reduced CDR are potential signs of glaucoma [75]. Thus, having systems to assess the presence of those lesions is relevant to assist the diagnosis as performed by the ophthalmologists. Physicians receive, in that case, not only a fundus image to analyze but also a report indicating where the pathological structures were detected, with a given probability of certainty [176]. Moreover, the automated quantification of the amount or severity of the lesions can help to grade the stage of the disease and detect the most urgent cases.

Several methods based on pattern recognition and machine learning have been developed to identify the presence of the lesions and to quantify their properties. The main advantage of this strategy is the comprehensibility of the results. Furthermore, as this approach reproduces the clinical practices that are usually followed by the physicians, ophthalmologists can assess their validity and decide whether the output of the system is correct or not [51].

The main disadvantage of these methods is that they are affected by the quality of the underlying segmentation approaches [24]. The automated recognition of each specific pathology requires to count with a dedicated method sufficiently robust to deal with the large variety of shapes, textures and color intensities of the lesions. This issue have motivated an increased effort in the literature to develop more accurate strategies for lesion detection, or to move through image characterization using global features.

We will analyze different systems for DR and glaucoma assessment based on lesion detection in the following sections. It must be mentioned, however, that the

Figure 2.12: Examples of red lesions as observed in fundus images. Notice that some of them are barely noticeable. Source: DIARETDB1 training set [97].

Figure 2.13: Examples of bright lesions as observed in fundus images. Source: DIARETDB1 training set [97].

segmentation methods used for detecting the lesions will not be discussed. Details about current state of the art methods for detecting red lesions will be provided in Chapter 5.

2.4.3.1 Diabetic retinopathy

Lesion-based systems for automated DR screening are mostly focused on detecting red lesions (MAs and small HEs) and bright lesions (exudates and cotton wool spots) [57]. The assessment of proliferative cases also requires the automated recognition of neovascularized areas [174].

Red lesions are challenging structures to detect due to their small size and the similarity of their intensities with respect to other objects such as the retinal vessels or the scars produced by laser treatments [180] (Figure 2.12). If images have low

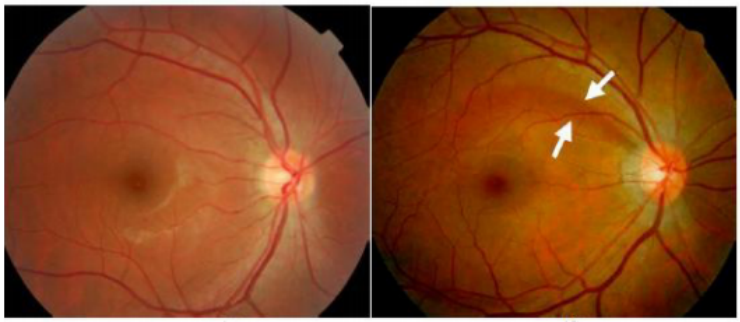
quality or contrast, it is difficult to distinguish them even with the naked eye. Bright lesions (Figure 2.13), on the contrary, are more easily to be recognized, but their identification usually requires to also separate the ONH to avoid false positives in that region [57]. Finally, a vessel characterization approach is usually needed to distinguish the areas where neovessels are occurring and to distinguish them the non-pathological vessels [174].

One of the earliest approaches for lesion-based detection of DR is the method presented in [68], which used neural networks for detecting exudates and hemorrhages, and to finally assess the existence of the disease. Since then, other screening systems were developed based on similar characteristics, but improving the quality of the segmentation strategies [168]. Other contributions were made in terms of the quantification of the lesions. The method presented in [139] and extensively evaluated in [176], for example, is based on the detection of red and bright lesions too, but the structures are quantified in terms of their area and probabilities, and this information is afterwards used to train a classifier that grades the images. A similar approach but based only on red lesions was recently presented in [179]. Automated PDR detection, on the other hand, requires to first detect neovascularizations, a task for which a series of different methods were recently introduced [110, 174].

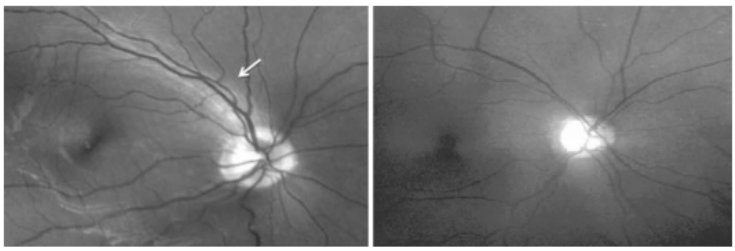
The extensive research in this field has motivated the translation of some of these systems to the clinical environment, with a significant success supported by several studies [219]. However, there is a consensus in that methods based on lesion detection still require to be improved, both developing better methods for detecting the structures involved or integrating further information from other sources such as medical records [186]. In [208], for example, the IRIS (Intelligent Retinal Imaging System) tool was compared with respect to the assessment performed by a team of experts in a retrospective cohort analysis. Although results were promising, the authors explicitly mentioned that the algorithms must be refined to achieve better performance [208].

2.4.3.2 Glaucoma

The detection of increased CDR or RNFL defects and peripapillary atrophy (PPA) is the most commonly used criteria for computer-assisted diagnosis of glaucoma [75].



(a) Localized RNFL defects. Left: healthy patient. Right: glaucomatous patient.



(b) Diffuse RNFL defects. Left: healthy patient. Right: glaucomatous patient.

Figure 2.14: Retinal Nerve Fiber Layer (RNFL) defects due to glaucoma, as manifested in fundus photographs. Source: [75].

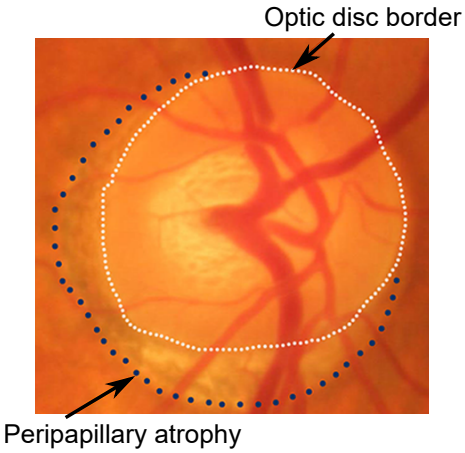


Figure 2.15: Peripapillary atrophy due to glaucoma. Source: [117].

RNFL (Figure 2.14) appears in fundus images as bright bundle striations that are unevenly distributed in normal eyes (Figure 2.14(a)). In glaucomatous patients,

however, localized defects in the RNFL such as those depicted in Figure 2.14(a) might appear [89]. Moreover, the loss of ganglion cells due to glaucoma can produce diffuse RNFL defects as those shown in Figure 2.14(b). The CDR is usually computed by first segmenting the OD and the optic cup. This task is challenging due to the lack of depth information in color fundus images, which is essential for a proper delineation of the optic cup. Finally, PPA (Figure 2.15) is observed in fundus images as a textured region surrounding the ONH.

Most of the literature in lesion-based glaucoma screening is based on the computation of the CDR value. There are numerous works trying to solve the task of segmenting the OD and the optic cup for CDR assessment. Although covering each of these strategies is not the purpose of this thesis, the interested reader could refer to the comprehensive review in [75] to know the most recent advances in the field. In general, the segmentation of the optic cup requires to use complementary features such as the curvature of the vessels to better discriminate the depth of the excavation on the ONH. In [91], for example, vessel bends are used as a shape prior to delineate the optic cup accurately. In other cases, vessels are inpainted in the image so their intensities do not interferes in the segmentation of the ONH [75].

In general, the other lesions associated with glaucoma are less evident than the increased in the CDR value. Thus, detecting RNFL defects or PPA using automated methods is significantly difficult [75]. However, a few efforts can be found in the literature. The method proposed in [133] is based on analyzing the textural properties of a vessel inpainted image to estimate the location of potential RNFL defects. Similarly, texture characteristics of the fundus are used in [146] for segmenting the RNFL defects using a Markov Random Field [101]. On the other hand, an automated system for PPA assessment on glaucoma cases was proposed in [117].

2.4.4 Image characterization using global features

Another family of methods relies on characterizing the images globally using overall image descriptors. This is achieved by extracting features that describes the distribution of the image intensities, first, and then training a classifier using those features as obtained from healthy and disease scans [75, 131]. This approach has the advantage of not requiring to segment any lesions or anatomical structures. Instead,

the image properties are expected to be sufficiently influenced by the changes associated with the diseases, allowing to ignore the segmentation step [75]. However, an important limitation is that these methods rely on hand-crafted features that require a significant engineering effort to develop. Moreover, ophthalmologists are not able to validate the features used to generate the diagnostic but only the diagnostic itself [75]. Their generalization error is usually higher, as well, as features tend not to be sufficiently robust to deal with the variety of situations that can be observed in a clinical setting [131]. More recently, deep learning based approaches (Section 2.4.5) have demonstrated to better deal with these last difficulties.

2.4.4.1 Diabetic retinopathy

A small number of methods for DR detection using overall features are available, compared to those based on lesion segmentation, for example. In [7], a strategy for characterizing patients with DR using the multiscale amplitude-modulation-frequency-modulation (AM-FM) method is presented. Instead of using all the image, authors sampled a set of windows around lesions that are typical from DR (such as MAs, HEs, neovascularization or exudates) and around normal regions (such as vessels or the retinal background). By adjusting a series of parameters, a set of feature vectors describing the texture of the retrieved patches is obtained. Then, the dimensionality of the feature vector is reduced using Principal Component Analysis (PCA), and a logistic regression classifier is applied to get the final label of the image. Similarly, descriptors obtained using higher order spectra (HOS) are used in [4] for characterizing the images, and a support vector machine (SVM) classifier is used to obtain a DR risk index. This strategy was further improved in [5] by modifying the features. Instead of using HOS, then, textures as characterized by gray level co-occurrence (GLCMs) and run-length (RLMs) matrices are applied for characterizing the images globally.

2.4.4.2 Glaucoma

Similarly to DR, only a few methods based on overall image features are available for glaucoma identification. In [26], features based on pixel intensities, textures, the coefficients of the Fast Fourier Transform (FFT) and a histogram model are

proposed for characterizing images with glaucoma. Instead of using the entire image as the information source, authors used a vessel inpainted version of the original color image that is centered in the ONH and downsampled to a lower resolution. Different classifiers and combinations of the features are used, and the performance is reported for the best configuration. Based on this prior study, a method for computing a glaucoma risk index was developed [27]. Such an approach modifies the previously published method by adding an illumination correction step to the preprocessing. Moreover, the selected features are both the pixel intensities, the FFT coefficients and the B-spline coefficients [27]. After a Principal Component Analysis (PCA) step that reduces the dimensionality of the feature vector, a two-stage classification step is performed in which a glaucoma score is obtained for each of the feature sources, and then used to train a separated classifier that determines the glaucoma risk index itself. More recently, Dua *et al.* [50] proposed to use a set of Wavelet-based energy features to characterize the texture of glaucomatous scans and discriminate them from healthy subjects. Four different classifiers were trained and evaluated using these features, achieving a high accuracy on a private data set. BRIEF (Binary Robust Independent Elementary Features) are also texture descriptors that were proposed in [130] for glaucoma identification, showing better performance than other characterization strategies when combined with a support vector machine (SVM) classifier.

To the best of our knowledge, CatEye [125] is the only automated system for glaucoma screening based on overall image features that was clinically evaluated. Its core classifier is based on the glaucoma risk index presented in [27]. According to [75], non-segmentation based might not be sufficient to be used in a clinical setting, as the features cannot be directly assessed by the ophthalmologists. However, authors propose to combine such strategies with other lesion-based approaches to improve the classification accuracy of currently available methods.

2.4.5 Deep learning based methods

Deep learning groups a variety of methods for classification and regression that are based on deep neural networks (DNNs) [25]. In particular, deep learning refers to the task of training a neural network to solve a given task. The term *deep* is associated

with the idea of having numerous layers of processing units, which resembles the organization of the neurons in the brain. Thus, given a large training set composed of a series of images and their annotations, DNNs are able to learn by themselves the features that are needed to solve the objective task. The process of learning itself consists on minimizing a cost function that penalizes any missclassification on the training set.

Although deep learning represents a cutting edge technology that is actively investigated by the research community, neural networks existed for decades [109]. However, the lack of large annotated data sets and the absence of high performance devices to perform the learning task make unfeasible to learn sufficiently deeper networks. As a consequence, networks with a shallow architecture were not able to improve their generalization error, which means that they were not able to scale to other real data sets different that the ones used for training [108]. However, recent advances in optimization techniques, Graphical Processing Units (GPUs) and parallel computing, and the existence of larger data sets or tools for large scale annotation such as Amazon Mechanical Turk [211] have renewed the interest in these methods [108]. Since the publication of Krizhevsky *et al.* [103], Convolutional Neural Networks (CNNs), which are a specific type of DNNs, have been used and modified to solve a large variety of tasks, including image classification [181] and semantic segmentation [182], for example.

The main advantage of deep learning techniques is their ability to automatically learn the best features to perform the classification task. This allows to avoid the tedious and time consuming process of feature design and selection, which usually involves decades of intensive research. In the particular case of automated detection of retinal diseases, DNNs are robust enough to even avoid the lesion segmentation step, as they can learn how to discriminate the diseases by recognizing the lesions as local features [73]. These features are usually difficult to be assessed by an expert, although recent studies have proposed strategies to represent them visually to solve this issue (Figure 2.16).

The precision of DNNs is achieved at the cost of learning from extremely, large annotated data sets. Thus, their applicability is restricted to domains that have large amounts of labelled data [87], which is normally difficult to obtain as image

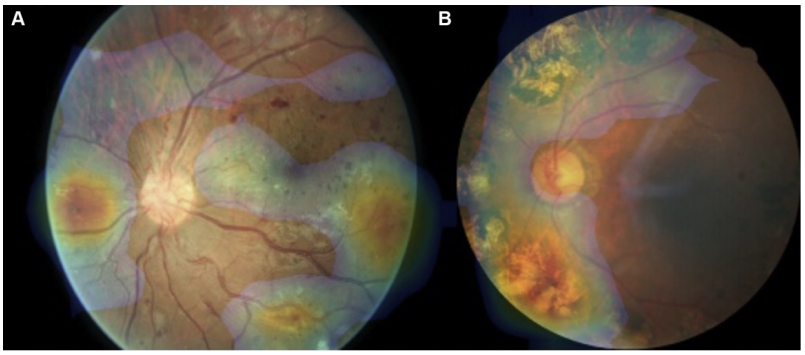


Figure 2.16: Abnormality heat maps as obtained by the deep learning based method for DR detection proposed by Gargeya *et al.* [69]. Source: [69].

acquisition and labelling require experts and costly devices [22]. This last characteristic makes the usage of CNN feasible only when large amounts of organized data are available. Moreover, the training process demands the availability of high performance computing hardware too, or it could take years to get a proper model or an architecture for the problem to solve.

2.4.5.1 Diabetic retinopathy

As mentioned above, the usage of deep learning techniques is limited by the availability of data. This made unfeasible the aplicability of DNNs for solving the automated DR detection task for years. However, the EyePACS data set [45] that was recently released as part of the Kaggle challenge on DR grading [94] has promoted the development of this type of techniques. In [163], a CNN architecture is proposed for detecting different grades of DR. This approach was evaluated on a subset of 5000 validation images extracted from the original Kaggle set, and demonstrated to be robust for the discrimination of the more evident grades. However, mild DR was not detected properly, likely because images are downsampled before feeding the network, and MAs are not identified as a consequence.

Since DR labels are strongly dependent on the observed lesions, a significant effort is being made for not only labelling the images but also to automatically indicate the regions that were considered relevant for the CNN to assign the label. By means of such an approach, ophthalmologists are able to determine whether the label is

correct or not. Quellec *et al.* [165], for instance, adapted CNNs so that different score maps representing the activation areas can be retrieved. Authors showed that by jointly optimizing both the classification accuracy and the produced heat maps, fuzzy segmentations of red and bright lesions were obtained. Similarly, the method presented in [69] and trained on the EyePACS data set is able to provide abnormality heat maps to highlight subregions of the images in which the model detects lesions (Figure 2.16). As an alternative, Wang *et al.* [212] proposed to add a regression activation map to a CNN to recover the most relevant areas.

2.4.5.2 Glaucoma

The development of automated methods for detecting glaucoma using deep learning is limited by the fact that no data sets with glaucoma annotations are currently publicly available. Instead, the vast majority of them contains labellings of the ONH, which make them useful for training OD and optic cup segmentation algorithms that may or not be based on deep learning [119]. To the best of our knowledge, the only existing work applying CNNs for automated detection of glaucoma is the one presented in [34]. In this paper, a CNN trained on images cropped around the ONH is able to achieve a significant performance when applied on two well known private data sets. As the model is trained on a relatively small data set, the architecture is not so deep. Moreover, no visualization strategy is proposed to validate the criteria followed by the CNN at the moment of labelling the images, so the learned features cannot be assessed by the experts.

Chapter 3

Blood vessel segmentation

Most of current systems for computer-aided diagnosis of eye diseases require first to automatically segment the retinal vasculature. This is motivated by the fact that blood vessels are, jointly with the optic nerve head (ONH), the most visible anatomical part of the retina. Moreover, it was demonstrated that a large variety of ophthalmic and system diseases are manifested as changes in the overall architecture and distribution of the retinal vessels.

In this chapter we propose a novel method for blood vessel segmentation in fundus photographs based on a fully connected conditional random field model. Standard segmentation priors such as a Potts model or total variation usually fail when dealing with thin and elongated structures such as vessels. We propose to overcome this difficulty by using a conditional random field model with more expressive potentials, taking advantage of recent results enabling inference of fully connected models almost in real-time. Parameters of the method are learned automatically by means of a structured output support vector machine, a supervised technique widely used for structured prediction in several machine learning applications. Our method, trained using state of the art features, is evaluated both quantitatively and qualitatively on publicly available data sets: DRIVE, STARE, CHASEDB1 and HRF. Additionally, a quantitative comparison with respect to other strategies is provided. Results show that this approach outperforms other techniques when evaluated in terms of standard quality metrics. Additionally, it was observed that the fully connected model is able to better distinguish the desired structures than the local neighborhood based

approach. Results suggest that this method is suitable for the task of segmenting elongated structures, a feature that can be exploited to contribute with other medical and biological applications.

This chapter is organized as follows: Section 3.1 summarizes the importance of blood vessel segmentation for fundus image analysis. Section 3.2 explains in detail our method. In Section 3.3 we provide information about the data sets and the quantitative measures used in our experiments. Section 3.4 presents our results, including a comparison to other recently published approaches. Section 3.5 discusses the advantages of the proposed method and further lines of research. Finally, Section 3.6 concludes the chapter.

An open source implementation of our approach is made publicly available in `https://github.com/ignaciorlando/fundus-vessel-segmentation-tbme`. The work presented in this chapter is published in:

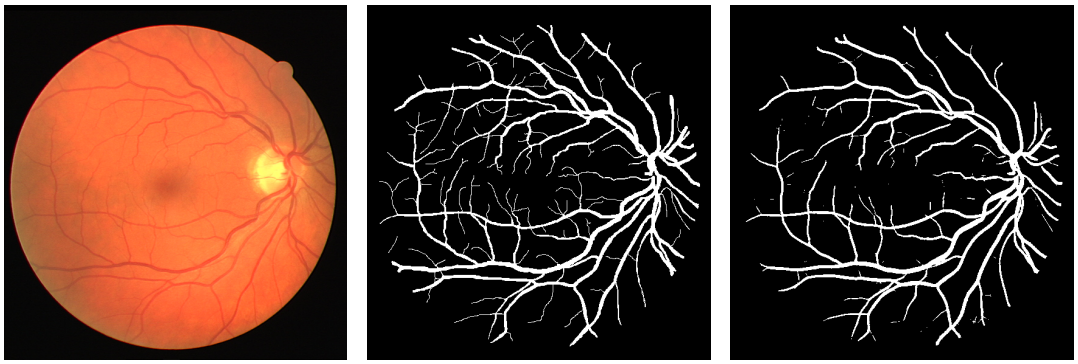
- J. I. Orlando and M. del Fresno. Reviewing preprocessing and feature extraction techniques for retinal blood vessel segmentation in fundus images. *Mecánica Computacional*, XXXIII(42):2729–2743, 2014

- J. I. Orlando and M. B. Blaschko. Learning fully-connected CRFs for blood vessel segmentation in retinal images. In P. Golland, C. Barillot, J. Hornegger, and R. Howe, eds., *MICCAI 2014, LNCS*, volume 8149, pp. 634–641. Springer, 2014

- J. I. Orlando, E. Prokofyeva, and M. B. Blaschko. A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images. *IEEE Transactions on Biomedical Engineering*, 64(1):16–27, 2017

3.1 Motivation

Retinal blood vessels are one of the most relevant anatomical landmarks in fundus photographs (Figure 3.1). They provide valuable information for the diagnosis, screening, treatment and evaluation of both DR and glaucoma [2]. As described in Chapter 2, the analysis of morphological attributes of retinal blood vessels allows the



(a) Fundus photograph (b) Manual annotations (c) Automated segmentation

Figure 3.1: A fundus photograph from the test set of DRIVE [141] and its vessel annotations as obtained by a human expert and the segmentation approach proposed in this thesis.

assessment of these diseases (see Section 2.4.2). In other cases, vessels need to be previously detected in order to facilitate the automation of the detection of lesions with similar intensities, such as microaneurysms and hemorrhages [52], or to obtain a more accurate segmentation of anatomical parts such as the optic disc or the optic cup [75]. The vascular bifurcations can also be used as landmarks for retinal image registration [2], which is valuable for the assessment of different images of the same patient that were acquired using different modalities, such as a fundus photograph and a fluorescein angiography.

Any automated analysis of the retinal vasculature requires its accurate segmentation first. In current best practice, this task is performed manually by trained experts, although this is particularly tedious and time-consuming. Furthermore, difficulties in the imaging process–such as inadequate contrast between vessels and background, and uneven background illumination–and the variability of vessel width, brightness and shape, reduce significantly the coincidence among segmentations performed by different human observers [46]. These facts motivate the development of automatic strategies for blood vessel segmentation without human intervention [2].

Although numerous attempts have been made in the field, this task is still an active area of research due to the potential impact of having more accurate results [63]. In general, existing approaches can be classified into two main categories, supervised and unsupervised. Supervised methods require a set of training samples–

typically composed of pixels features and their known annotations–to learn a model or a classifier. Several classifiers have been considered in the literature, including k-nearest neighbors [141], Bayesian [188], support vector machines [220, 225], neural networks [121, 201], decision trees [64, 65], Gaussian mixture models [46], AdaBoost [118], among others. A trainable filter, named B-COSFIRE, was recently introduced in [20] to highlight the retinal vasculature. Though the method is not supervised in the sense of training a classifier, the strategy they follow to adjust its parameters is based on training data. Convolutional Neural Networks (CNNs) have also been applied for retinal blood vessel segmentation more recently [67, 119]. By contrast, unsupervised methods are systems that are able to segment the vasculature without requiring any manual annotations, although typically at the cost of lower accuracy. In general, most of these strategies are based on applying thresholding, vessel tracking techniques [223] or region-oriented approaches–such as region growing [61, 62, 173] or active contours [8, 228]–after vessel enhancement. This last task is performed by means of morphological operations [226], matched filter responses [31, 95, 145], the complex continuous wavelet transform [56], among others [190]. The method we propose in this thesis belongs to the supervised category.

Conditional Random Fields (CRFs) are extensively used for image segmentation in several applications [79, 104, 113]. To the best of our knowledge, however, they were never applied before to blood vessel segmentation in fundus images. This is likely due to that the standard pairwise potentials, such as in a Potts model, assign a low prior to the elongated structures that comprise a vessel segmentation. This fact motivated us to introduce a novel method for blood vessel segmentation based on fully connected CRFs. These models were previously applied in [92] and [93] for liver and brain tumor segmentation in CT and MRI, but their implications on the segmentation of two dimensional, thin structures was not previously studied. In our experiments we demonstrate that the dense connectivity augments the capability of the method to detect elongated structures, overcoming the original difficulty of local neighborhood based CRFs and improving results significantly. This property can potentially contribute to a number of different biological and medical applications where the segmentation of such structures is required, including automatic plant root phenotyping [60] or neuron analysis [80].

As is shown in [102], local classification leads to misclassification issues that might arise while incorporating prior knowledge about the shape of the desired structures on the learning process. CRFs are able to provide such information through the pairwise potentials. Structured Output SVM (SOSVM) has been used before to learn local neighborhood based CRFs [88, 197]. However, learning dense CRFs using SOSVMs was avoided before due to its computational intractability, since the learning method requires multiple calls to the inference algorithm during training, and the inference in dense CRFs is usually slow. We overcome this problem by making use of recent advances in efficient inference in fully connected CRFs [102].

In this chapter, we present our method for blood vessel segmentation based on learning fully connected CRFs using SOSVMs. We also propose a strategy to estimate additional parameters of the method such as the configuration of features in order to optimize its performance during training. The method is evaluated both quantitatively and qualitatively on four standard and publicly available data sets (DRIVE, STARE, CHASEDB1 and HRF) to study the behavior of the algorithm under different contexts, including images of healthy patients, containing pathologies and taken at different resolutions. According to our experiments, this method outperforms current strategies when evaluating in terms of several different quality measures.

3.2 Methods

This section explains in detail our method. First, both the local neighborhood based and the fully connected CRF formulations are described (Section 3.2.1). Afterwards, we summarize the strategy to learn such models by means of a SOSVM (Section 3.2.2). The features used to evaluate our method are explained in Section 3.2.3. Finally, Section 3.2.4 describes a compensation factor that can be used to segment images with different resolutions without needing to recalibrate feature parameters.

3.2.1 Conditional Random Fields for vessel segmentation

The segmentation task can be posed as an energy minimization problem in a conditional random field (CRF). In the original definition of CRFs, images are mapped to graphs, where each pixel represents a node, and every node is connected with an edge to their neighbors according to a certain connectivity rule [102, 104, 105]. In local neighborhood based CRFs, nodes are connected following a 4 pixel neighborhood connectivity [28], while in the fully connected definition each node is assumed to be linked to every other pixel of the image [102].

We denote by $\mathbf{y} = \{y_i\}$ a labeling over all pixels of the image I in the label space $\mathcal{L} = \{-1, 1\}$, where 1 is associated to blood vessels and -1 to any other class. A conditional random field $(I, \mathbf{y})$ is characterized by the Gibbs distribution:

$$p\left(\mathbf{y}|I\right) = \frac{1}{Z(I)} \exp\left(-\sum_{c \in \mathcal{C}_{\mathcal{G}}} \Phi_c(\mathbf{y}_c|I)\right) \tag{3.1}$$

where $Z(I)$ is a normalization constant, $\mathcal{G}$ is the graph associated to I and $\mathcal{C}_{\mathcal{G}}$ is a set of cliques in $\mathcal{G}$, each inducing a potential Φ_c [113]. This distribution states the conditional probability of a labeling $\mathbf{y}$ given the image I. The Gibbs energy function can be derived from this likelihood:

$$E(\mathbf{y}|I) = \sum_{c \in \mathcal{C}_{\mathcal{G}}} \Phi_c(\mathbf{y}_c|I) \tag{3.2}$$

Thus, the maximum a posteriori (MAP) labeling can be obtained by minimizing the corresponding energy:

$$\mathbf{y}^* = \arg\min_{\mathbf{y} \in \mathcal{L}} E(\mathbf{y}|I) \tag{3.3}$$

After minimizing $E(\mathbf{y}|I)$, a binary segmentation of the vasculature is obtained. For notational convenience, we will omit the conditioning in the rest of this chapter, and we will use $\psi_c(\mathbf{y}_c)$ to denote $\Phi_c(\mathbf{y}_c|I)$. Additionally, we will consider energies that decompose as summations over unary and pairwise potentials, in contrast to more general higher order potentials [101].

Given a graph $\mathcal{G}$ on $\mathbf{y}$, its energy is obtained by summing its unary and pairwise

potentials (ψ_u and ψ_p, respectively):

$$E(\mathbf{y}) = \sum_i \psi_u\left(y_i, \mathbf{x}_i\right) + \sum_{(i,j)\in\mathcal{C}_\mathcal{G}} \psi_p\left(y_i, y_j, \mathbf{f}_i, \mathbf{f}_j\right) \tag{3.4}$$

where $\mathbf{x}_i$ and $\mathbf{f}_i$ are the unary and pairwise features, respectively. Unary potentials define a log-likelihood over the label assignment $\mathbf{y}$, and they are traditionally computed by a classifier [102]. Pairwise potentials define a similar distribution but considering only the interactions between pixels features and their labels, according to $\mathcal{C}_\mathcal{G}$, which is determined by the graph connectivity.

Unary potentials are common to both the local neighborhood based and the fully connected CRF, and they are obtained as follows:

$$\psi_u\left(y_i, \mathbf{x}_i\right) = -\langle \mathbf{w}_{u_{y_i}}, \mathbf{x}_i \rangle - \mathbf{w}_{\beta_{y_i}}\beta \tag{3.5}$$

where β is a bias constant, and $\mathbf{w}_{u_{y_i}}$ and $\mathbf{w}_{\beta_{y_i}}$ represents the weight vectors for the features and the bias term, respectively, both associated to the label y_i. The unary vector $\mathbf{x}_i$ is given by an arbitrary combination of features extracted from the image.

Pairwise potentials are defined as a linear combination of functions. Thus, our pairwise energy is obtained according to:

$$\psi_p\left(y_i, y_j, \mathbf{f}_i, \mathbf{f}_j\right) = \mu(y_i, y_j) \sum_{m=1}^{M} {w_p}^{(m)} k^{(m)}\left(f_i^{(m)}, f_j^{(m)}\right) \tag{3.6}$$

where each $k^{(m)}$ is a fixed function over an arbitrary feature $f^{(m)}$, ${w_p}^{(m)}$ is a linear combination weight, and $\mu(y_i, y_j)$ represents a label compatibility function. The Gaussian kernels determine the similarity between connected pixels by means of $f^{(m)}$. Since the neighboring information is provided by the connectivity rule followed by the model, these kernels depend on the CRF formulation, so they are described afterwards. The remaining terms are detailed in the sequel.

The compatibility function μ penalizes similar pixels that are assigned to different labels, and it is given by the Potts model $\mu(y_i, y_j) = [y_i \neq y_j]$, where Iverson bracket notation $[\cdot]$ indicates one if the statement is true and zero otherwise.

Parameters $\mathbf{w}_u$, ${\mathbf{w}_p}^{(m)}$ control the relevance of the unary features and the pairwise

kernels on the energy function, respectively. Additionally, $\mathbf{w}_\beta$ is used to learn the bias term. The adjustment of these parameters is not feasible to be done manually due to their high dimensionality, so we propose to learn them using a Structured Output SVM, as is explained in detail in Section 3.2.2.

3.2.1.1 Local neighborhood based CRFs

Local neighborhood based CRFs (LNB-CRFs) are defined over grid graphs. Thus, in this type of model each node (pixel) is assumed to be connected by an edge to its 4-connected neighbors. The function for the pairwise potentials given the m-th pairwise feature is obtained as follows:

$$k^{(m)}\left(f_i^{(m)}, f_j^{(m)}\right) = \frac{|f_i^{(m)} - f_j^{(m)}|}{2\theta_{(m)}^2} \tag{3.7}$$

where $\theta_{(m)}$ is a bandwidth that controls the relevance of the dissimilarities between pixel features. The energy of the grid based model is minimized using the min-cut/max-flow approach proposed by [28].

3.2.1.2 Fully connected CRFs

In a fully connected CRF model (FC-CRF), each node of the graph is assumed to be linked to every other pixel of the image. Using these higher order potentials, the method is able to take into account not only neighboring information but also long-range interactions between pixels. This property improves the segmentation accuracy, but makes implementation of the inference process computationally expensive in general. Recently, however, Krähenbühl and Koltun [102] have introduced an efficient inference approach under the restriction that the pairwise potentials are a linear combination of Gaussian kernels over an Euclidean feature space. This approach, which is based on taking a mean field approximation of the original CRF, is able to produce accurate segmentations in a few seconds.

Pairwise kernels for the fully connected model have the following form:

$$k^{(m)}\left(f_i^{(m)}, f_j^{(m)}\right) = \exp\left(-\frac{|\mathbf{p}_i - \mathbf{p}_j|^2}{2\theta_p^2} - \frac{|f_i^{(m)} - f_j^{(m)}|^2}{2\theta_{(m)}^2}\right) \tag{3.8}$$

where $\mathbf{p_i}$ and $\mathbf{p_j}$ are the coordinate vectors of pixels i and j. Positions are included in the pairwise terms to increase the effect of close pixels over distant ones. Kernel widths θ_p and $\theta_{(m)}$ control the degree of relevance of the two parts of the kernels in the expression. For instance, when θ_p increases, much longer interactions are taking into account. On the contrary, when θ_p decreases, only local neighborhoods significantly affect the result. Similarly, when $\theta_{(m)}$ increases or decreases, higher or lower differences on the m-th feature are tolerated, respectively.

3.2.2 Learning CRFs with Structured Output SVM

Our goal is to learn a vector $\mathbf{w} = (\mathbf{w}_u, \mathbf{w}_\beta, \mathbf{w}_p)$, where $\mathbf{w}_u$, $\mathbf{w}_\beta$ and $\mathbf{w}_p$ are the weights for the unary features, for the bias term and for the pairwise kernels, respectively. The vector $\mathbf{w}$ can be high-dimensional if multiple features are considered, so manual or automated adjustment using techniques such as grid search is not feasible in a reasonable time. Supervised learning of the unary potentials separately from the pairwise potentials might be an alternative, but this approach ignores the influence of the pairwise potentials on the general energy formulation, and can lead to worse results than joint learning of the weights. We therefore propose to obtain $\mathbf{w}$ in a supervised way, using the 1-slack formulation of the SOSVM with margin-rescaling presented in [88]. Such a discriminative training approach has shown promising results for building highly complex and accurate models in several areas, including object detection, image segmentation and computer vision applications, even for large datasets. To the best of our knowledge, however, it was never used before for the task of learning FC-CRFs.

Let the training set $S = \{\left(s^{(1)}, y^{(1)}\right), ..., \left(s^{(n)}, y^{(n)}\right)\}$, where n is the number of training images. Each $y^{(i)}$ corresponds to the ground truth of the i-th image in the training set. Each set $s^{(i)} = \{x^{(i)}, \beta, f^{(i)}\}$ contains the set $x^{(i)}$ of unary feature vectors, a bias constant $\beta = 1$, and the set $f^{(i)}$ of pairwise features for every pixel in

the image.

The weights $\mathbf{w}$ are obtained by solving:

$$\min_{\mathbf{w},\xi\geq 0} \frac{1}{2}\|\mathbf{w}\|^2 + C\xi \tag{3.9}$$

subject to

$$\forall \left(\bar{y}^{(1)}, ..., \bar{y}^{(n)}\right) : \sum_{i=1}^{n} \langle \mathbf{w}, \psi(s^{(i)}, y^{(i)}) - \psi(s^{(i)}, \bar{y}^{(i)})\rangle \geq \sum_{i=1}^{n} \Delta(y^{(i)}, \bar{y}^{(i)}) - \xi \tag{3.10}$$

where C is a regularization constant; ξ is a slack variable shared across all the constraints $\bar{y}^{(i)}$; $\varphi(s, y)$ is a feature map function that relates a given set s with a given labelling y; and $\Delta(y, \bar{y})$ is a loss function that evaluates the difference between a ground truth y and a constraint $\bar{y}$. In this work, we define Δ as the Hamming loss:

$$\Delta(y, \bar{y}) = \sum_{i} [y_i \neq \bar{y}_i] \tag{3.11}$$

where we used Iverson bracket notation. This function penalizes all the differences between the predicted labelling and the gold standard segmentation.

Our feature map is defined as follows:

$$\varphi(s, y) = \left(\sum_{k} \varphi_u(\mathbf{x}_k, y_k), \sum_{k} \varphi_\beta(\beta, y_k), \sum_{k} \sum_{j<k} \varphi_p\left(y_k, y_j, \mathbf{f}_k, \mathbf{f}_j\right) \right) \tag{3.12}$$

where the components represent the sum of the unary feature map, the bias feature map and the pairwise feature map, respectively, for all the pixels in the image. We give precise definitions of φ_u, φ_β, and φ_p in the sequel.

We define a binary vector $\varphi_y(y_i) \in \{0, 1\}^{|\mathcal{L}|}$ such that:

$$\varphi_y(y_i) = \begin{cases} (1, 0) & if \quad y_i = -1 \\ (0, 1) & if \quad y_i = 1 \end{cases} \tag{3.13}$$

The individual feature maps are obtained as follows:

$$\varphi_u(\mathbf{x}_k, y_k) = \mathbf{x}_k \otimes \varphi_y(y_k) \tag{3.14}$$

$$\varphi_\beta(\beta, y_i) = \beta\varphi_y(y_i) \tag{3.15}$$

$$\forall m : [\varphi_p(y_k, y_j, \mathbf{f}_k, \mathbf{f}_j)]_m = \mu(y_i, y_j)k^{(m)}(f_i^{(m)}, f_j^{(m)}) \tag{3.16}$$

where $\otimes$ is the Kronecker product. We solve Equation (3.9) efficiently using the cutting-plane approach proposed in [88].

3.2.3 Features

We evaluated our method using features that are widely used in the field of blood vessel segmentation in fundus images: responses to the multiscale line detectors presented by Nguyen *et al.* [138] and responses to 2D Gabor wavelets [188] are used to compute the unary potentials, and a vessel-enhanced image processed with the method by Zana and Klein [226] for the pairwise potentials. An extensive analysis of other features was performed in [152].

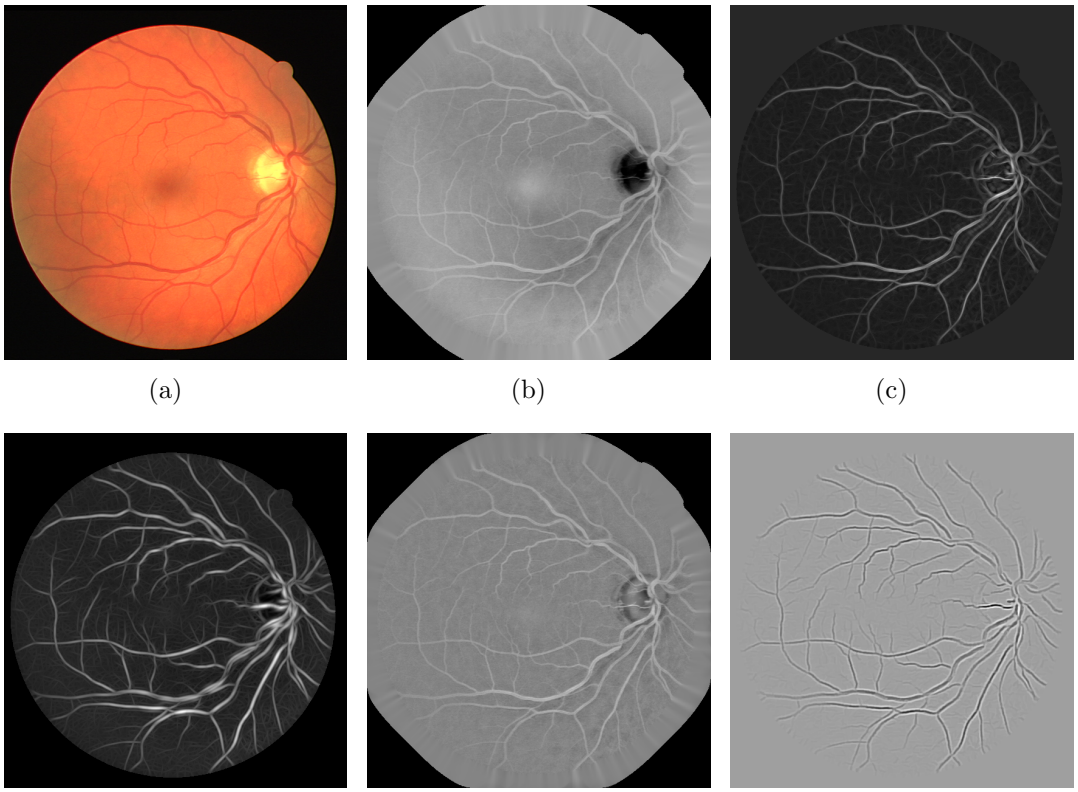
All features are extracted from grey scale images, obtained by taking the inverted green band of the original, RGB color image, as reported by other works [20, 121]. Additionally, due to false detections introduced by the selected features on the border of the FOV, we replicate the strategy proposed in [188] to simulate a wider aperture of the capture device. By means of this technique, false detections occurring outside the original FOV can be easily removed by multiplying the resulting image with the original FOV mask. An example of the resulting preprocessed image is shown in Fig. 3.2(b).

Nguyen *et al.* line detectors exploit the property that blood vessels appear as elongated structures. The average intensity is calculated along a line of length l passing through each target pixel $\mathbf{P}$ at different orientation angles α. The line with the largest mean intensity $L_l(\mathbf{P})$ is selected from all the considered orientations, and the line strength of the pixel is computed by taking the difference $S_l(\mathbf{P}) = L_l(\mathbf{P}) - N_s(\mathbf{P})$, with $N_s(\mathbf{P})$ being the average intensity in a square window centered on $\mathbf{P}$ with edge length s. An example of the responses obtained with $l = 15$ is shown in Fig. 3.2(c). The original version of this feature combines responses at different scales and the inverted green channel into a single feature, which is then thresholded.

Here we take each S_l and the inverted image separately, since our method is able to learn the best weights to combine the features. Thus, instead of having a single value per pixel, we have a feature vector composed of the responses to each value of l and the image I.

2D Gabor wavelets have the capability to detect oriented features and can be tuned to specific frequencies. This property is especially useful to enhance the vasculature, since blood vessels appear at different sizes and orientations. We compute this feature exactly as reported by Soares *et al.* [188] at different scales a. Responses of the image to this wavelet, taken at different values of a, are included as features. Fig. 3.2(d) depicts an example obtained with $a = 3$.

Zana and Klein's technique for vessel enhancement takes advantage of the fact that the vessels are linear, connected, and their curvature varies smoothly along the crest line [226]. Noise of the image is first reduced by applying an opening by reconstruction operation, using linear structuring elements of length l at different angles. Afterwards, multiple top-hat morphological operations are applied using the same structuring elements, and the sum of the corresponding responses for each given angle is taken. This transformation reduces small bright noise and improves the contrast of all linear components. Structures whose curvature is linearly coherent are then detected by means of a cross-curvature evaluation, performed by applying a Laplacian of Gaussian with windows of size 7×7 pixels and standard deviation 1.75. Finally, an alternating filter composed of a successive application of a morphological opening, a closing and an opening is applied to remove false detections of non linear patterns on bright or dark thin irregular zones and background linear features. In the three last operations, the same linear structuring element of length l is used. We have observed that this feature is highly sensitive to uneven illumination of the fundus, degrading its ability to characterize the blood vessels effectively. In order to improve its quality we incorporated an additional preprocessing, only for this feature, where an estimated background is subtracted from the green band of the original color image. The background is estimated by convolving the green band with a median filter, where the size of the filter kernel is large enough to ensure that the blurred image contains no visible structures such as vessels. This approach has been applied several times in the literature [121, 126], and an example of the



(a) (b) (c)

(d) (e) (f)

Figure 3.2: Image preprocessing and unary and pairwise features examples. (a) Original color image. (b) Inverted green band after border expansion. (c) Response to Nguyen *et al.* line detector ($l = 15$). (d) Response to Soares *et al.* 2D Gabor wavelet at the scale $a = 3$. (e) Inverted image after bias correction. (f) Image enhanced using Zana and Klein method ($l = 9$).

resulting image is illustrated in Fig. 3.2(e). The Zana and Klein feature is illustrated in Fig. 3.2(f).

All features are normalized independently to zero mean and unit variance, using the mean and the standard deviation of each feature calculated on each image [188].

3.2.4 Scaling Models to Images of Different Resolution

Although the weights for the unary features and the pairwise kernels are adjusted during the learning process, system performance is still related to the capability of

the features to effectively characterize vascular structure. In general, features are sensitive to their parameters, which are usually related to vessel properties such as their calibre, which is at the same time related to the image resolution. Responses to the 2D Gabor wavelet, for example, depend on the scale a. Similarly, Nguyen *et al.* line detectors and the Zana and Klein enhancement strategy depend on the length l of the detectors or the linear structuring element, respectively. Most of these feature parameters were originally set using low resolution images, such as those in the DRIVE dataset [193]. When applying such features to higher resolution images, performance is significantly reduced if the feature extraction procedure is not proportionately scaled. Other parameters such as the angles for computing feature responses at different orientations are not influenced by changes in the resolution of the images.

A similar behavior can be expected for preprocessing parameters, e.g. the size of the median filter used to estimate the background, or the size of the aperture simulated by the border expansion. The parameter θ_p used on the pairwise potentials of the FC-CRF is also influenced by image resolution, since it weighs the pairwise interactions according to the relative distance of each pixel.

The proper adjustment of such parameters is relevant when applying the framework to images of different resolution, and having an automatic strategy for their calibration is valuable. Grid search using labelled images in the training set is computationally prohibitive due to the high dimensional space that comprises the parameters and their combinations. As an alternative to manual adjustment, some authors propose to derive parameters related to the vessel calibre from the width of the vessel of interest [20] or the size of the optic nerve head [160]. However, both methods require prior knowledge about structures that are difficult to measure and vary from one image to another. We introduce in the sequel a different strategy to automatically adapt features and model parameters to images of different resolution.

Instead of adjusting the configuration for each single data set resolution, we propose a simple approach based on estimating the best configuration of feature parameters on a single data set, and then adapting such parameters by multiplying them with a compensation factor $\rho = \frac{\mathcal{X}_{\text{new}}}{\mathcal{X}_{\text{training}}}$, where $\mathcal{X}_{\text{training}}$ represents the average width of the FOV in the images used to configure the scales, and $\mathcal{X}_{\text{new}}$ is the average

width of the FOV in the new images. As changes in the resolution are expected to be related to changes in the number of pixels associated to the FOV region, this approach approximates invariance of feature computation with respect to scaling.

3.3 Materials and Evaluation

This section describe the data sets and the metrics used to evaluate our method. Additionally, we provide further details about the strategy followed to estimate the parameters θ_p and $\theta_{(m)}$ of the CRF, and the C parameter of the SOSVM (Eq. (3.9)).

3.3.1 Datasets

Our experiments were carried out on DRIVE [193], STARE [83], CHASEDB1 [64] and HRF [144, 145], four standard, publicly available data sets of fundus images used for the evaluation of blood vessel segmentation algorithms.

DRIVE[1] includes 40 color fundus photographs (7 of them with pathologies) with a 45° FOV, with 8 bits per color channel at 565×584 pixel resolution. The set is divided into a training and a test set, both containing 20 images. Two different manual annotations are provided for the test set, and only one annotation per image is available on the training set.

STARE[2] comprises 20 images, 10 of them containing pathologies, captured at 35° FOV, with 8 bits per color channel and a resolution of 700×605 pixels. Two observers manually segmented all images, with the second observer marking vessels with thinner annotations than the first one [46]. Despite this variability in annotation methodology between the two observers, performance is normally evaluated using the first observer's segmentation as the ground truth [20, 121]. FOV masks are not provided in the original set, so the masks built by Marin *et al.*[3] [121] were used. The set is not divided into training and test, and no consistent evaluation methodology has emerged from the reviewed literature. In order to be able to compare our method with respect to the largest amount of the available literature we performed our

[1] `http://www.isi.uu.nl/Research/Databases/DRIVE/`
[2] `http://www.ces.clemson.edu/~ahoover/stare/probing/index.html`
[3] `http://www.uhu.es/retinopathy/muestras2.php`

evaluation on STARE using leave-one-out cross-validation.

CHASEDB1[4] contains images of each eye of 14 children, comprising a total of 28 images. Pictures were captured with 30° FOV, using 8 bits per color channel at 1280×960 pixels, centered on the optic disc. Two expert labelings per image are provided. FOV masks were obtained using the approach proposed in [20], as they are not provided in the original dataset. The 28 images are divided into training and test, with 8 and 20 images in each set, respectively [64]. The first 20 images are used for testing, and the last 8 images for training.

The HRF[5] data set contains 15 images of healthy patients, 15 images of patients with diabetic retinopathy and 15 images of glaucomatous patients. Images were captured with 60° FOV and 3304×2336 pixel resolution. Only one ground truth segmentation per image is available, generated by a group of experts. To the best of our knowledge, this data set was not previously employed to evaluate supervised vessel segmentation algorithms, so we constructed a training set comprising the first 5 images of each subset, and tested on all remaining images. To reduce the computational cost of our experiments, images and labels on the training set and images on the test set were downsampled by a factor of 2, and results were afterwards upsampled so they can be compared with respect to the original manual annotations.

3.3.2 Evaluation Metrics

Results were analyzed quantitatively by comparing our segmentations with the gold standard labelings provided on each data set. Seven different measurements were obtained, all of them in terms of the number of true positives TP, true negatives TN, false positives FP and false negatives FN, and considering only the pixels inside the FOV:

$$Se = \frac{TP}{TP + FN}, \quad Sp = \frac{TN}{TN + FP}, \quad Pr = \frac{TP}{TP + FP}$$

$$F1 = \frac{2 \cdot Pr \cdot Re}{Pr + Re}, \quad G = \sqrt{Se \times Sp}$$

[4] `http://blogs.kingston.ac.uk/retinal/chasedb1/`

[5] `https://www5.cs.fau.de/research/data/fundus-images/`

$$MCC = \frac{TP/N - S \times P}{\sqrt{P \times S \times (1-S) \times (1-P)}}$$

where $N = TP + TN + FP + FN$ is the total number of pixels of the image, $S = (TP + FN)/N$ and $P = (TP + FP)/N$. Sensitivity (Se, also known as Recall Re) measures the capability of the method to properly detect blood vessels, while specificity (Sp) is an indicator of the capability of distinguishing all other non-vessel structures. Sp suffers in the presence of imbalanced classes due to the low influence of the false positives term in the denominator of the fraction. By contrast, Se does not present this issue as it involves a fraction of pixels corresponding to the vessel class. However, though a higher Se value is desired, it must be analyzed in combination with Sp, as Se can be trivially maximized by labeling all pixels as vascular. Precision (Pr) quantifies the ratio of pixels classified as vessel that are correctly identified. The Accuracy (Acc) is not included as it is sensitive to unbalanced distributions in the number of pixels belonging to the positive and negative classes [115].We include the Matthews Correlation Coefficient (MCC), the F1-score (F1) and the G-mean (G), which are overall performance measures that are more suitable to imbalanced class ratios. The MCC is a correlation coefficient between the manual and predicted binary segmentations, and it has been previously used for the evaluation of retinal vessel segmentation methods [20, 61, 201]. It returns a value between 1 and +1, with +1 indicating a perfect prediction, 0 no better than random, and 1 a total disagreement between prediction and ground truth. The F1-score is the harmonic mean of precision and recall, and it also has the property of better characterizing quality when data are imbalanced. It achieves its maximum value of 1 when the segmentation of the positive class is perfect, and its lowest value of 0 when the segmentation is completely wrong. Similarly, the G-mean is a metric that measures the balanced between Se and Sp by taking their geometric mean, returning a value between 0 and 1 [78]. Finally, receiver operating characteristic (ROC) curves were also generated from the unary potentials and the energy of the FC-CRF, and the area under each curve (AUC) was computed. The AUC on STARE was obtained by computing the ROC curve on each image and taking the average value.

3.3.3 Model selection

Parameters for computing the unary features were initially fixed to the values reported by the original references, which use the DRIVE training set [138, 188]. Thus, responses to the 2D Gabor wavelet were obtained at scales $a = 2, 3, 4, 5$, and responses to line detectors were analyzed from $l = 1$ to 15, with increments of $l_0 = 2$. As Zana and Klein used different data to estimate the length of the structuring element, we selected $l = 9$, which is consistent with the average calibre of DRIVE vessels as reported in [10]. The size of the windows used for preprocessing the image to compute this feature was fixed at 35 pixels, and the border of the FOV was expanded by 50 pixels. For datasets other than DRIVE, we made use of the compensation factor ρ described in Section 3.2.4. In the case of the Nguyen line detector, the increment l_0 is also multiplied to reduce the dimensionality of the feature vector when evaluating on images with higher resolutions. The angles were set to the values reported in the original references [138, 188, 226].

To estimate the parameter C of the SOSVM we randomly divided each training set into two new subsets, *training** and *validation*, containing 70% and 30% of the *training* images, respectively [77]. We used *training** to train the model, and *validation* to estimate the best C value. A model selection phase was initially performed, in which the SOSVM was trained using different values of C. Performance of each trained model was evaluated on the validation set. Values of $C \in \{(10^i)/c\}$, with $i \in \{-2, ..., 0, ..., 3\}$ with c equal to the total number of FOV pixels in the training set were evaluated. The configuration was selected to maximize the average F1-score on the validation set. That learned configuration of the CRF model was then evaluated on the test set only once. This configuration prevent us from using any test data to adjust parameters, allowing us to obtain a non-biased estimate of the test error [77]. Cross-validation was only used on the STARE dataset due to its limited size and the absence of a test set.

A similar approach was followed with the purpose of adjusting θ_p: for a fixed value of C, different θ_p values spanning from 1 to 15 were explored, and the one that maximized the F1-score on the validation set was chosen. This search, however, was performed only once, using the DRIVE validation set. At the time of evaluation on the remaining sets, the selected value $\theta_p = 5$ was multiplied by the compensation

Table 3.1: Evolution of F1-score during forward feature selection on DRIVE.

Features	**Unary potentials**			**Pairwise potentials**	
	Iter. 1	**Iter. 2**	**Iter. 3**	**Iter. 1**	**Iter. 2**
Line detector	0.6898	**0.7256**	-	0.7535	0.6985
2D Gabor wavelet	**0.6967**	-	-	0.7423	0.7437
Zana and Klein enhanced	0.6378	0.7043	0.7129	**0.7546**	-

factor.

Forward feature selection [116] using DRIVE's *training** and *validation* sets was followed to identify which feature combinations are more suitable for the unary and the pairwise potentials. Table 3.1 shows the progress in the mean F1-score obtained on the validation set for each configuration of features in each iteration. Numbers in bold indicate that the set of features was chosen in that iteration. We observed that the 2D Gabor wavelet contributes to detect thicker structures but with a large number of false positives. By adding responses to the Nguyen *et al.* line detector, false positives are reduced and narrower vessels are segmented. Pairwise potentials showed better performance when using the Zana and Klein vessel enhanced image in the kernel. This configuration was applied to all subsequent data sets.

A strategy to estimate the scale parameter of a radial basis function is to take the median of all pairwise distances of a random sample of pixels [178]. Since part of each pairwise kernel consists of a radial basis function (Eq. 3.6), this approach was applied to the estimation of $\theta_{(m)}$. This estimator is robust in that it has low variance when it is computed over different random samples [178]. However, small changes to $\theta_{(m)}$ can affect the results due to the exponentiation in the pairwise term and the number of interactions taking into account by the fully connected model. Therefore, we estimate $\theta_{(m)}$ as the median of medians obtained over 50 different random samples of pairs $(f_i^{(m)}, f_j^{(m)})$ extracted from the training set of each data set.

3.4 Experiments and Results

In this section we present the results obtained in our experiments. The prototype of this method was implemented with MATLAB R2013a, using MEX functions to interface with C++ implementations of the LNB-CRF and the FC-CRF. In Sec-

tion 3.4.1 we summarize and analyze the results, while in Section 3.4.2 we include a comparison with respect to other published methods.

3.4.1 Results

A quantitative evaluation of the results obtained on our experiments using only the unary potentials and using the FC-CRF is presented in Table 3.2. The binary segmentations were obtained by minimizing the corresponding energies using the mean field approximation strategy proposed in [102]. Results obtained with the LNB-CRF are not included in the table as they are exactly the same than those obtained using only the unary potentials. This is due to the fact that, in this configuration, the SOSVM assigns an almost-zero value to $\mathbf{w}_p$, the parameter that weights the local pairwise potentials, and at the same time it does not modify the weights associated to the unary potentials with respect to the configuration achieved when only the unary potentials are considered. This is because the grid connectivity does not provide valuable information in the context of elongated structures such as retinal vessels, in line with our original hypothesis. By contrast, SOSVM assigns a non-zero value to $\mathbf{w}_p$ when training the FC-CRF, and also modifies the weights associated to the unary features and the bias term, meaning that the pairwise potentials influence the other parameters and contribute substantially to the prediction function. To evaluate the statistical significance of such influence, a set of right-tailed Wilcoxon signed-ranks hypothesis tests were performed on the quality values obtained using only the unary potentials and using the FC-CRF. No hypothesis tests were performed on STARE results since segmentations on this set were obtained by leave-one-out cross validation. Taking into account that the parameter C was tuned according to a validation set subsampled from each training set, it is not possible to assume that all the results were achieved with the same configuration. In addition to the quantitative analysis, we also provide several segmentation examples to analyze qualitatively the changes introduced by the pairwise potentials.

In some of the data sets the FC-CRF contributes to a statistically significant improvement in the results when evaluating in terms of F1-score (DRIVE: $p \approx 4 \times 10^{-5}$; CHASEDB1: $p \approx 1 \times 10^{-3}$), G-mean (DRIVE: $p \approx 4 \times 10^{-5}$; CHASEDB1: $p \approx 8 \times 10^{-5}$; HRF: $p \approx 9 \times 10^{-7}$) and MCC (DRIVE: $p \approx 4 \times 10^{-5}$; CHASEDB1:

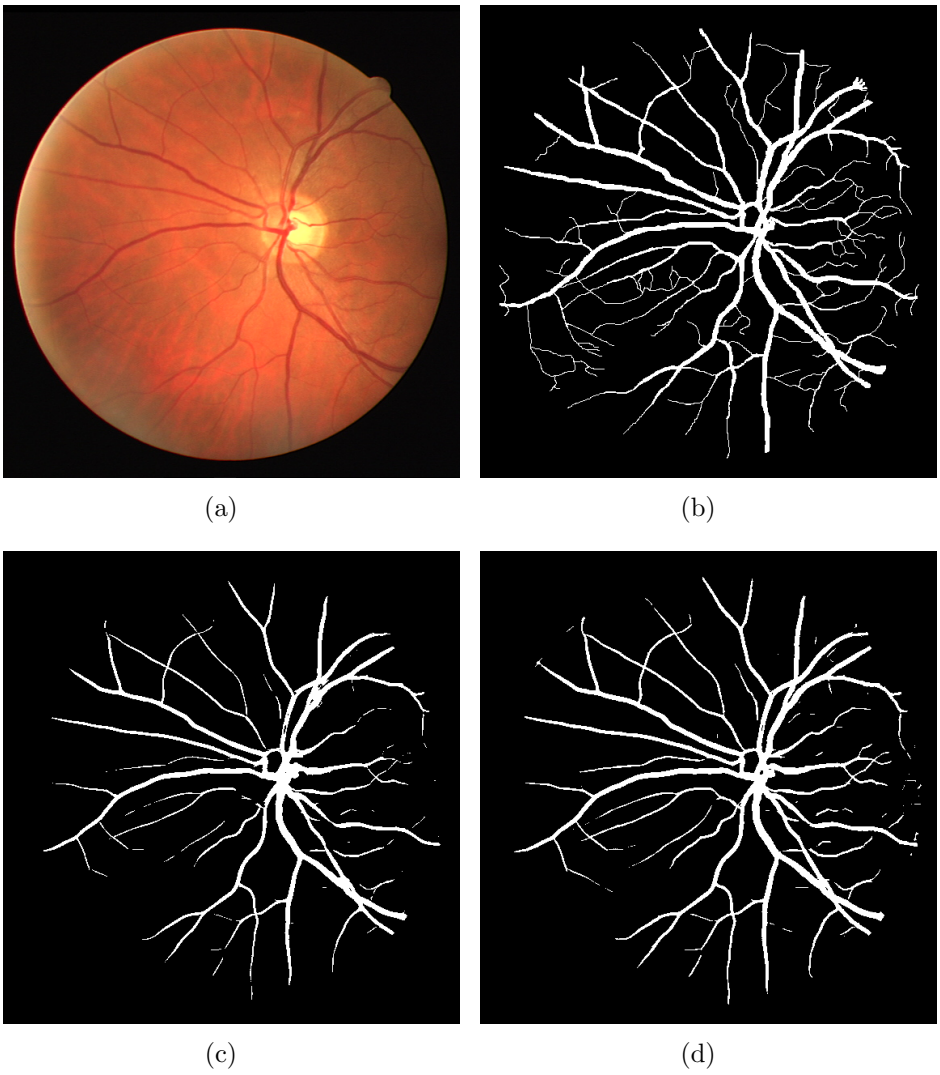
Table 3.2: Quantitative evaluation of the results obtained on DRIVE, STARE, CHASEDB1 and HRF, using only the unary potentials (UP) or the fully connected CRF (FC-CRF).

Dataset	Method	Se	Sp	Pr	F1	G	MCC
DRIVE	UP	0.7079	**0.9802**	**0.8394**	0.7661	0.8324	0.7401
	FC-CRF	**0.7897**	0.9684	0.7854	**0.7857**	**0.8741**	**0.7556**
STARE	UP	**0.7692**	0.9675	0.7445	0.7517	0.8618	0.7252
	FC-CRF	0.7680	**0.9738**	**0.7740**	**0.7644**	**0.8628**	**0.7417**
CHASEDB1	UP	0.7110	0.9707	0.7386	0.7209	0.8304	0.6919
	FC-CRF	**0.7277**	**0.9712**	**0.7438**	**0.7332**	**0.8403**	**0.7046**
HRF	UP	0.7315	**0.9680**	**0.7012**	0.7127	0.8413	0.6851
	FC-CRF	**0.7874**	0.9584	0.6630	**0.7158**	**0.8686**	**0.6897**

$p \approx 4 \times 10^{-3}$; HRF: $p \approx 2 \times 10^{-2}$). F1-score and MCC are improved on average on STARE and HRF, as indicated in Table 3.2. In the case of HRF, the improvement in the F1-score is lower than that obtained on other data sets. G-mean is also increased on average on STARE.

When evaluating on DRIVE, the pairwise potentials improve the Se value ($p \approx 4 \times 10^{-5}$) and slightly reduce the average Sp ($p \approx 4 \times 10^{-5}$). This is likely due to the FC-CRF introducing a certain fraction of false positives, as it can be observed in the reduction of the average Pr ($p \approx 4 \times 10^{-5}$). In all such cases, however, the fraction of improvement in Se is higher than the reduction on Sp and Pr, which is evidenced by the improvement in the G-mean, and also in the F1-score and the MCC values. Some examples of the segmentations obtained on DRIVE are shown in Fig. 3.3. It is possible to observe that the FC-CRF model incorporates a number of thinner vessels and significantly improves the connectivity of the vascular structure.

In STARE, the G-mean is slightly improved when using the dense approach. When decomposed into their terms, it is possible to see that the Se is slightly reduced in average, but with an improvement in both the average Sp and Pr, which is associated with a reduction in the number of false positive pixels. An example of this setting is observed in Fig. 3.5, which depicts an extreme pathological case. The FC-CRF contributed to reducing the number of false positives in the haemorrhage inside the optic disc. Yet the second human observer identified vessels within that region, the first human observer (which is assumed as the ground truth) did not



(a) (b)

(c) (d)

Figure 3.3: Segmentation results obtained on DRIVE. (a) Image 04 of DRIVE. (b) Ground truth labelling. (c) Segmentation obtained using only the unary potentials. (d) Segmentation obtained using the FC-CRF.

mark anything there, which directly affects the Se value. Some narrow structures are integrated to the vascular tree by the FC-CRF model, and it can also be observed that the unary potentials overestimated the width of some of the major vessels.

In CHASEDB1, results are increased in terms of both Se ($p \approx 7 \times 10^{-5}$) and Sp, although the improvement in this last metric is not statistically significant. As a consequence, the G-mean is increased. The average Pr value is also improved by

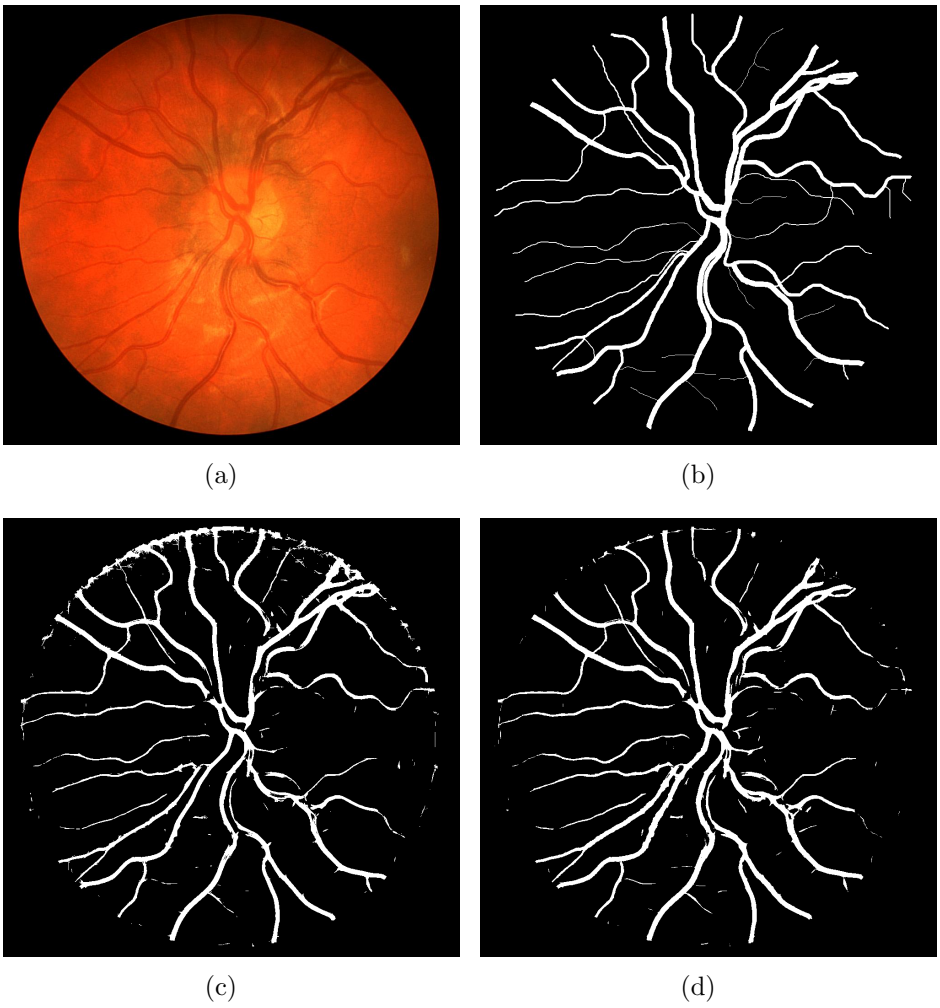
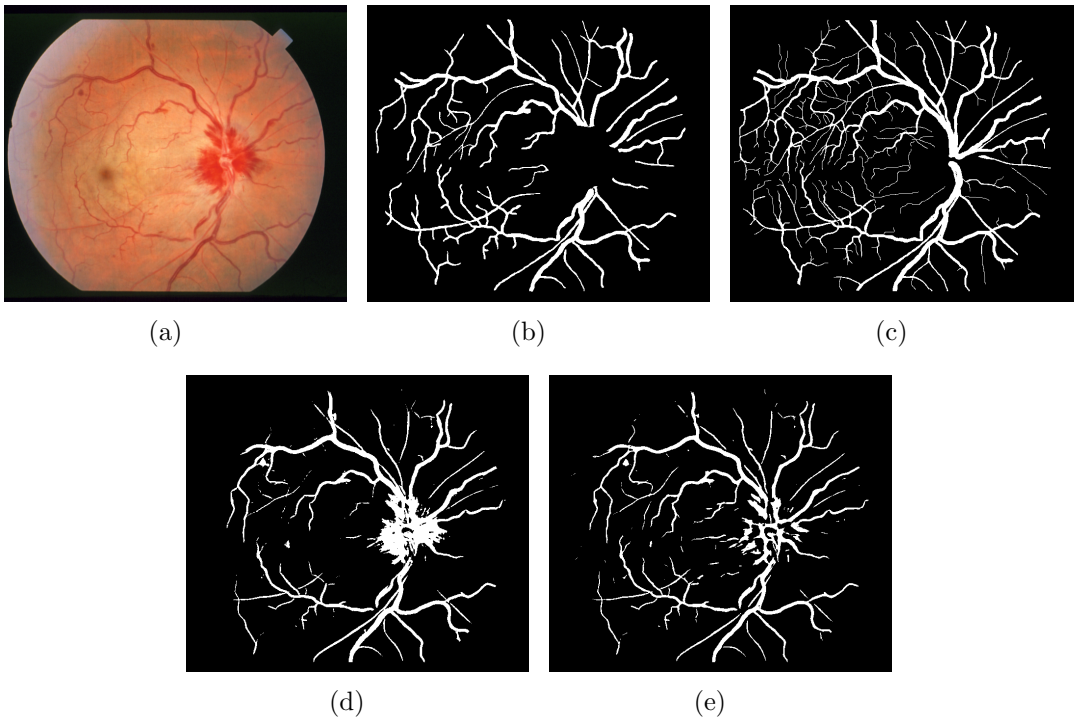


Figure 3.4: Segmentation results obtained on CHASEDB1. (a) Image_05L of CHASEDB1. (b) Ground truth labelling. (c) Segmentation obtained using only the unary potentials. (d) Segmentation obtained using the FC-CRF.

the FC-CRF, which is explained by the reduction in the number of false positives, as seen in Fig. 3.4. It is also possible to observe that in both test sets, the unary potentials overestimate the calibre of the narrower vessels, a setting that is improved when incorporating the pairwise potentials.

A different behavior can be observed on HRF (Fig. 3.6), where Se is significantly increased ($p \approx 9 \times 10^{-7}$) but Sp ($p \approx 9 \times 10^{-7}$) is diminished. Pr is also decreased ($p \approx 9 \times 10^{-7}$), meaning that a number of false positives in introduced. We observed



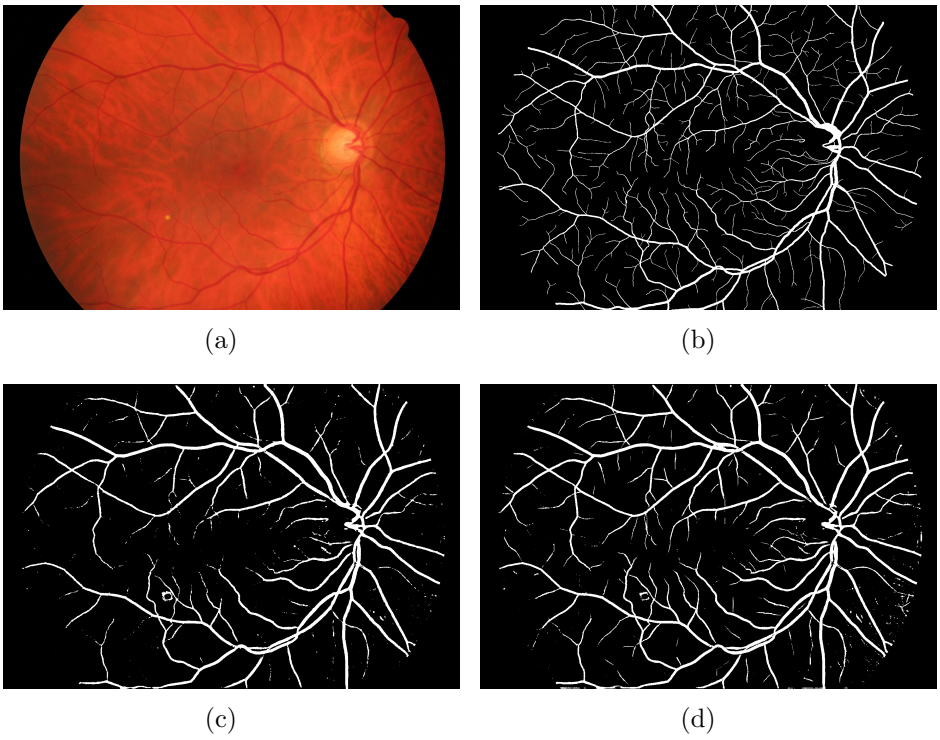
(a) (b) (c)

(d) (e)

Figure 3.5: Segmentation results obtained on a serious pathological case on STARE. (a) Image im0005. (b) First human observer annotations. (c) Second human observer annotations. (d) Segmentation obtained using only the unary potentials. (e) Segmentations obtained using the FC-CRF model.

qualitatively, however, that the FC-CRF detects a large number of narrow vessels that are ignored when using only the unary potentials, as can be seen in Fig. 3.7. It is possible to see also that some of the thinner capillaries remain ignored. Additional work on feature construction might help to incorporate such structures. It can be seen that the pairwise potentials contribute to joining isolated detections, resulting in a more general connected tree, and in an increase in the G-mean value.

A comparison between the ROC curves obtained using only the unary potentials and using the FC-CRF can be observed in Fig. 3.8, where the second human observer performance on each set (if available) is plotted. The curve on STARE is not included since results there were obtained by cross-validation, and the segmentation of each image was made with a different model. Results on DRIVE show that the FC-CRF outperforms the unary potentials, and also that they are quantitatively tied to the



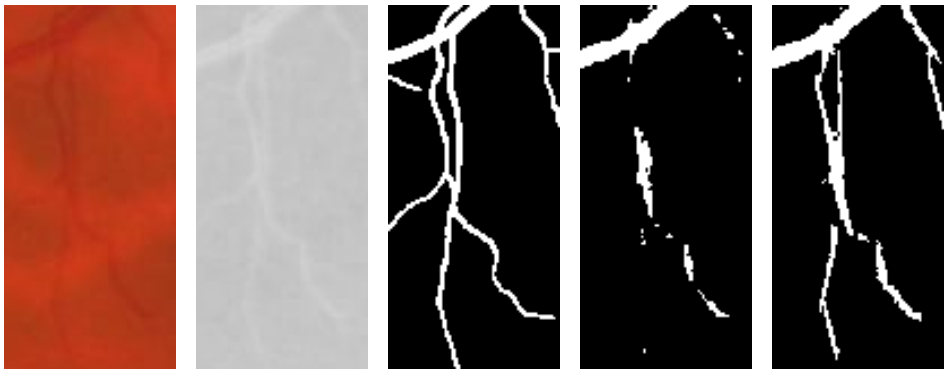
(a) (b)

(c) (d)

Figure 3.6: Segmentation results obtained on HRF test set. (a) Image 11_g. (b) Manual annotation. (c) Segmentation obtained using only the unary potentials. (d) Segmentation obtained using the FC-CRF.

second observer performance. When evaluating on HRF, however, and in line with the analysis previously performed, the unary potentials results in a better ROC curve than the FC-CRF. The areas under each of the curves are in line with these conclusions.

Finally, the computation cost of our single-thread, single-core implementation of the inference on the FC-CRF model was evaluated on an Intel(R) Xeon(R) CPU E5-2690 0 platform at 2.90GHz with 64 GB of RAM. For this purpose, the average time of applying the FC-CRF on each test set was measured. As seen in Fig 3.9, although the computational cost grows with the resolution of the images, it is still fast enough to be feasibly applied in a clinical setting.



(a) (b) (c) (d) (e)

Figure 3.7: Example of narrow vessel detection under low contrast conditions. (a) Detail of Image 11_g. (b) Preprocessed image. (c) Manual annotation. (d) Segmentation obtained using only the unary potentials. (e) Segmentation obtained using the FC-CRF.

3.4.2 Comparison with other methods

We also include a comparison of our results with respect to those reported by other state-of-the-art methods evaluated on DRIVE (Table 3.3) and STARE (Table 3.4), CHASEDB1 and HRF (Table 3.5). Although our method is supervised, we also compare with unsupervised approaches in the tables. Methods that obtained the final binary segmentations using parameters that were estimated on the test data were not included on the tables of DRIVE and STARE, since that setting might underestimate the actual test error [74, Section 6].[6] Results obtained similarly but on CHASEDB1 were included in Table 3.5 as they are the only works evaluated on those sets. However, they are marked with an asterisk.

Our method reports the highest F1-score and MCC on DRIVE when compared with other supervised and unsupervised strategies. Se is also higher, for a relatively acceptable Sp value. As indicated previously, the Sp measure describes the capability of the method to distinguish the non-vessel class, and it usually suffers when the

[6] In [20], the results reported on DRIVE test set, STARE and CHASEDB1 correspond to the binary segmentations obtained by thresholding the responses to the B-COSFIRE filter. However, the threshold is selected by maximizing the average MCC on each corresponding test set. Similarly, in [64, 65, 121, 172] the scores provided by different classifiers are thresholded using the parameter that maximizes the average Acc on the test data (DRIVE test set, STARE and CHASEDB1 in the case of [64, 65, 172], and DRIVE test set and STARE in [121]).

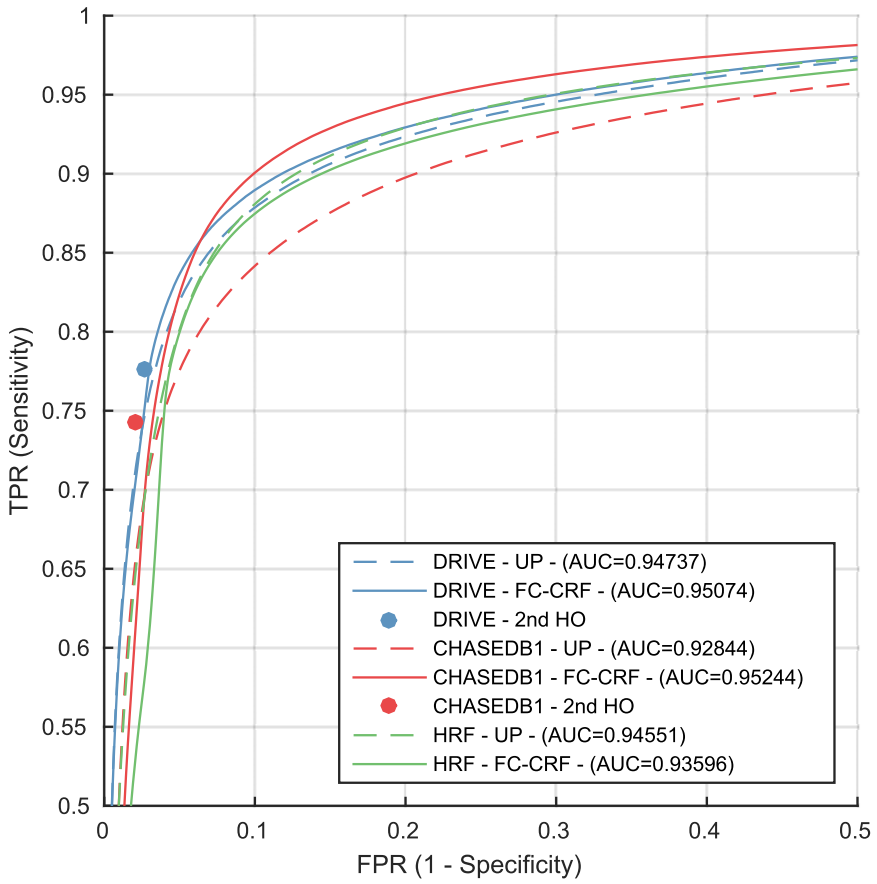


Figure 3.8: ROC curves on DRIVE, CHASEDB1 and HRF, using only the unary potentials (UP, slashed line) or the FC-CRF (solid line), and second human observer (HO) performance.

segmentations have a high number of false positives. However, the Pr value is higher than the method by Fathi *et al.* [56], which has reported a higher Sp value but a lower average Se.

Comparison on STARE is difficult as most of the state-of-the-art methods performed their analysis using their own strategies to train and test.[7] It must be taken into account, then, that the supervised methods listed in the comparison illustrated in Table 3.4 were trained miscellaneously. When evaluated on this set, our method

[7]Some methods were trained on the first half and tested on the second half of STARE [12] or even on the entire set [20], or were trained on a random sample of pixels extracted from STARE [188], etc.

Table 3.3: Comparison of average Se, Sp, Pr, $F1$-score, G-mean and MCC values of our method with respect to other existing blood vessel segmentation algorithms and the 2nd human observer, when evaluating on DRIVE.

	DRIVE					
Methods	**Se**	**Sp**	**Pr**	**F1**	**G**	**MCC**
FC-CRF	**0.7897**	0.9684	0.7854	**0.7857**	**0.8741**	**0.7556**
2nd human observer	0.7760	0.9730	0.8066	0.7881	0.8689	0.7601
Supervised						
Cheng *et al.* [36]	0.7252	0.9798	-	-	0.8429	-
Dai *et al.* [46]	0.7359	0.9720	-	-	0.8458	-
Niemeijer *et al.* [141]	0.6793	0.9725	-	-	0.8128	-
Lupascu *et al.* [118]	0.6728	**0.9874**	-	-	0.8151	-
Orlando and Blaschko [151]	0.7850	0.9670	0.7770	0.7810	0.8713	0.7482
Soares *et al.* [188]	0.7283	0.9788	-	-	0.8443	-
Xu and Luo [220]	0.7760	-	-	-	-	-
You *et al.* [225]	0.7410	0.9751	-	-	0.8500	-
Vega *et al.* [201]	0.7444	0.9600	-	0.6884	0.8454	0.6617
Unsupervised						
Amin *et al.* [12]	0.6608	0.9799	-	-	0.8047	-
Al-Diri *et al.* [8]	0.7282	0.9551	-	-	0.8340	-
Bankhead *et al.* [21]	0.7027	0.9717	-	-	0.8263	-
Budai *et al.* [30]	0.6440	0.9870	-	-	0.7973	-
Chakraborti *et al.* [31]	0.7205	0.9579	-	-	0.8308	-
Espona *et al.* [53]	0.6634	0.9682	-	-	0.8014	-
Espona *et al.* [53]	0.7436	0.9615	-	-	0.8456	-
Fathi and Naghsh-Nilchi [56]	0.7768	0.9759	0.7559	0.7669	0.8707	-
Fraz *et al.* [61]	0.7152	0.9768	**0.8205**	0.7642	0.8358	0.7333
Fraz *et al.* [62]	0.7302	0.9742	0.8112	0.7686	0.8434	0.7359
Martínez *et al.* [123]	0.7246	0.9655	-	-	0.8364	-
Miri *et al.* [128]	0.7352	0.9795	-	-	0.8486	-
Odstrcilik *et al.* [145]	0.7060	0.9693	-	-	0.8272	-
Palomera *et al.* [156]	0.6440	0.9670	-	-	0.7891	-
Roychowdhury *et al.* [173]	0.7390	0.9780	-	-	0.8501	-
Vlachos and Dermatas [204]	0.7468	0.9551	-	-	0.8446	-
Wang *et al.* [210]	0.7520	0.9800	-	-	0.8585	-
Yin *et al.* [223]	0.6522	0.9710	-	-	0.7958	-
Zhang *et al.* [227]	0.7120	0.9724	-	-	0.8321	-
Zhao *et al.* [228]	0.7420	0.9820	-	-	0.8536	-

Table 3.4: Comparison of average Se, Sp, Pr, $F1$-score, G-mean and MCC values of our method with respect to other existing blood vessel segmentation algorithms and the 2nd human observer, when evaluating on STARE.

	STARE					
Methods	**Se**	**Sp**	**Pr**	**F1**	**G**	**MCC**
FC-CRF	0.7680	0.9738	**0.7740**	**0.7644**	0.8628	**0.7417**
2nd human observer	0.8951	0.9387	0.6424	0.7401	0.9166	0.7225
Supervised						
Cheng *et al.* [36]	0.7813	**0.9843**	-	-	0.8769	-
Dai *et al.* [46]	0.7769	0.9550	-	-	0.8614	-
Soares *et al.* [188]	0.7200	0.9750	-	-	0.8379	-
You *et al.* [225]	0.7260	0.9751	-	-	0.8414	-
Vega *et al.* [201]	0.7019	0.9671	-	0.6082	0.8239	0.5927
Unsupervised						
Al-Diri *et al.* [8]	0.7521	0.9681	-	-	0.8533	-
Budai *et al.* [30]	0.5800	0.9820	-	-	0.7547	-
Chakraborti *et al.* [31]	0.6786	0.9586	-	-	0.8065	-
Fathi and Naghsh-Nilchi [56]	**0.8061**	0.9717	0.7027	0.7509	**0.8850**	-
Fraz *et al.* [61]	0.7409	0.9665	0.7363	0.7386	0.8462	0.7003
Fraz *et al.* [62]	0.7318	0.9660	0.7294	0.7306	0.8408	0.6908
Martínez *et al.* [123]	0.7506	0.9569	-	-	0.8475	-
Palomera *et al.* [156]	0.7790	0.9409	-	-	0.8561	-
Odstrcilik *et al.* [145]	0.7847	0.9512	-	-	0.8639	-
Roychowdhury *et al.* [173]	0.7320	0.9840	-	-	0.8487	-
Wang *et al.* [210]	0.7800	0.9780	-	-	0.8734	-
Yin *et al.* [223]	0.7248	0.9666	-	-	0.8370	-
Zhang *et al.* [227]	0.7177	0.9753	-	-	0.8366	-
Zhao *et al.* [228]	0.7800	**0.9780**	-	-	0.8734	-

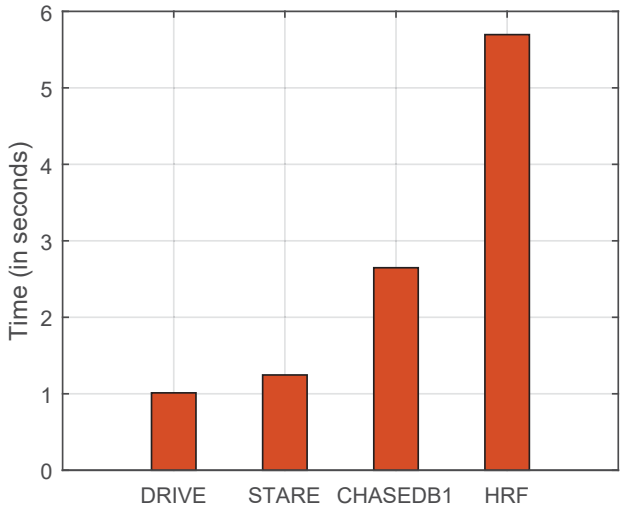


Figure 3.9: Computational cost of the FC-CRF inference in all the data sets used for evaluation.

reports the highest average F1-score and MCC values, indicating a better overall performance. Additionally, the FC-CRF outperforms the other strategies in terms of the Pr measure, meaning that the number of false positive detections is lower than in the other cases.

When evaluating on CHASEDB1 it is possible to see that our method achieved a better Se value than the other strategies. The F1-score value is outperformed, yet the results reported in the reference were obtained using parameters estimated from the test labels, and the parameters of the features it used were adjusted to this specific data set. In contrast, our FC-CRF model was trained using feature parameters that were scaled using the compensation factor.

To the best of our knowledge, only unsupervised methods were previously tested on HRF [30, 145]. We include the results of [145] on the test set calculated from the binary segmentations provided by the corresponding authors. In general, it is possible to see that the FC-CRF gives higher Se values than the method proposed in [145], but with lower Sp and Pr values. This means that the FC-CRF obtains a larger number of false positive detections than the other strategy. However, global metrics including F1-score and MCC are competitive.

Table 3.5: Comparison of average Se, Sp, Pr, $F1$-score, G-mean and MCC values of our method with respect to other existing blood vessel segmentation algorithms and the 2nd human observer, when evaluating on the CHASEDB1 and HRF.

CHASEDB1	**Se**	**Sp**	**Pr**	**F1**	**G**	**MCC**
FC-CRF	0.7277	0.9712	0.7438	0.7332	0.8407	0.7046
2nd human observer	0.7425	0.9793	0.8090	0.7686	0.8527	0.7475
Fraz *et al.* [65]*	0.7259	0.9770	0.7732	0.7488	0.8421	-
HRF	**Se**	**Sp**	**Pr**	**F1**	**G**	**MCC**
FC-CRF	**0.7874**	0.9584	0.6630	0.7158	**0.8687**	0.6897
Odstrcilik *et al.* [145]	0.7794	**0.9650**	**0.6950**	**0.7324**	0.8672	**0.7065**

3.5 Discussion

The FC-CRF model and the learning strategy better exploited the interaction between pixels features than the local neighborhood based approach. The local neighborhood approach was not able to improve results with respect to the unary potentials, as a zero weight is assigned to the pairwise term by the structured output SVM. The hypothesis tests performed on the results obtained by both the FC-CRF and the unary potentials on a number of different data sets, as explained in Section 3.4.1, demonstrated that the dense pairwise potentials introduced statistically significant improvements in several metrics. Additionally, as evidenced in ROC curves in Fig 3.8, the FC-CRF model also yields results that are tied to the second human observer performance. Such properties are due to the contribution of the high order pairwise potentials, which are able to better reconstruct the vessels even under low contrast conditions (Fig. 3.7) with a negligible time overhead (Fig 3.9). Although no directional prior is explicitly learned by the model, the combination of the distance and the feature dissimilarity terms within the pairwise kernel (Eq. (3.8)) provides a way to penalize too long or dissimilar interactions, respectively. Thus, if the pairwise features are robust enough, then the model will assign low energies to the labellings of filamentary structures, and will penalize other non elongated shapes. The features analyzed in this work consistently achieved better results than using only unary potentials, as observed in Table 3.1. By contrast, the LNB-CRF model is not able to take advantage of the pairwise features, as evidenced by the absense of improvement with respect to the unary potentials. Based on this property, it is possible to con-

clude that dense potentials are able to better characterize the vessels. Other medical and biological applications might benefit by using this approach for segmenting other tubular and elongated structures such as vessels, neurons or plant roots. For this last application, we have presented an unsupervised method based on FC-CRFs that was able to achieve promising results for segmenting *Arabidopsis thaliana* roots [153].

Extensive comparison with state of the art methods has also shown that our approach consistently performed well on several metrics, and is a fully-automated segmentation algorithm that achieves better results when evaluated in terms of global binary classification measures such as F1-score, G-mean and MCC. This is in part due to previous studies focusing on raw pixel accuracy, which ignores the fact that the number of pixels occupied by blood vessels is a relatively small fraction of the image. As a result, competing methods suffer as measured by F1-score, G-mean and MCC, which are particularly important as they reflect an accurate estimation of the vessel pixels, the primary goal in vessel segmentation for fundus image analysis.

As in the case of other supervised techniques–such as Gaussian mixture models [188] or SVMs [220]–the performance of our method is affected by the general ability of the features to characterize the retinal vasculature. State of the art features were used in our experiments in order to evaluate the contribution of the fully connected model in the improvement of the original results. Most of the features presented in the literature were designed using low resolution images such as those in DRIVE and STARE. Since the design of features involves the adjustment of different parameters that are effected by image scale, a decrease in performance is expected when the resolution of the images differs from the original setting. An alternative to reduce this effect would be to design the features for each specific resolution, though this is particularly time consuming. Recently, Vostatek *et al.* [206] proposed a different method for automatically adjust the parameters of the features to different image resolutions, based on linear regression. However, such a strategy requires to optimize the parameters for different resolution first, so that the line is adjusted according to these measurements. Instead, we proposed to use a simple technique based on applying a compensation factor ρ that is multiplied to each scale parameter before feature extraction. Using this basic approach our method is able to partially compensate for changes in resolution, outperforming other well-known segmentation strategies.

Although our approach achieves overall good performance, we have observed some missclassifications in the bright central reflex of the major arteries in the high resolution images in CHASEDB1 and HRF. This is likely due to the limited capability of both the unary and the pairwise features to deal with this property, as it was not taken into account when they were originally designed. Additional feature development or learning would be valuable in order to improve performance under this or other challenging contexts, such as in the presence of serious pathological changes. Currently, most features work well on lower resolution images, as fewer features have been developed for high resolution images such as HRF. It is a promising avenue of research to consider image structures that become apparent at higher resolutions, such as the central reflex in arteries. We encourage further research in this direction. Furthermore, evaluating other types of higher order potentials, combined with this learning approach, could potentially improve the results by capturing other types of pixel interactions. However, it must be taken into account that those approaches always involve a trade-off between performance and computational tractability.

3.6 Conclusions

In this chapter, we have presented a detailed description and evaluation of our discriminatively trained segmentation model based on a fully connected CRF for the purpose of blood vessel segmentation in fundus images. By means of features extracted from the images and fully connected pairwise potentials, this approach is able to reconstruct the retinal vasculature much more precisely than using only the unary potentials or a local neighborhood based CRF. The effectiveness of the approach is evidenced by the general improvement in the values of F1-score, G-mean and Matthews correlation coefficient–three quantitative measures that are suitable in binary classification problems where the number of true positive and true negative pixels are unbalanced–obtained on a number of benchmark data sets. ROC curves also show that results achieved with the FC-CRF are comparable with those obtained by a second human observer. The capability of the dense potentials to reconstruct elongated structures can potentially benefit other biological and medical applications.

Chapter 4

Transfer learning for glaucoma screening

Most current systems for automated glaucoma detection in fundus images rely on segmentation-based features, which are known to be influenced by the underlying segmentation methods. Convolutional Neural Networks (CNNs) are powerful tools for solving image classification tasks as they are able to learn highly discriminative features from raw pixel intensities. However, their applicability to medical image analysis is limited by the non-availability of large sets of annotated data required for training. In this chapter we present results of analysis of the viability of using CNNs that are pre-trained from non-medical data for automated glaucoma detection. Two different CNNs, namely OverFeat and VGG-S, were applied to fundus images to generate feature vectors. Preprocessing techniques such as vessel inpainting, contrast-limited adaptive histogram equalization (CLAHE) or cropping around the optic nerve head (ONH) area were explored within this framework to evaluate the improvement in feature discrimination, combined with both ℓ_1 and ℓ_2 regularized logistic regression models. Results on the DRISHTI-GS1 dataset, evaluated in terms of area under the average ROC curve, suggests the viability of this approach and offer significant evidence of the importance of well-chosen image pre-processing for transfer learning when the amount of data is not sufficient for fine-tuning the network.

This chapter is organized as follows. Section 4.1 provides insights about the

motivation to use transfer learning for glaucoma detection, including a review of current state of the art methods for performing this task. Section 4.2 details the strategy we developed, while Section 4.3 summarizes the CNN architectures and the data sets used in our experiments. In Section 4.4 we present the results obtained. Finally, Section 4.5 concludes the chapter.

An open source implementation of our approach is made publicly available in `https://github.com/ignaciorlando/overfeat-glaucoma`. The work presented in this chapter is published in:

- J. I. Orlando, E. Prokofyeva, M. del Fresno, and M. B. Blaschko. Convolutional neural network transfer for automated glaucoma identification. volume 10160, pp. 101600U–101600U–10. 2017. doi: 10.1117/12.2255740. URL `http://dx.doi.org/10.1117/12.2255740`

4.1 Motivation

Most recent strategies for automated screening of glaucoma are based on analysing properties of the ONH [91, 99], blood vessels [98, 218] or the nerve fiber layer [133, 146], for which it is essential to previously detect and segment these regions. As segmentation-based features are known to be significantly influenced by the precision of the underlying segmentation methods [75], an increasing scientific effort in the field of glaucoma detection is devoted to the development of new strategies based on overall image properties. As previously mentioned in Chapter 2, such approaches are based on transformations of the image intensities [27], texture analysis [130] or combinations of image derived features with clinical and genetic information [221]. An important limitation of these strategies, however, is that they rely on hand-crafted features, which require a significant engineering effort to develop [32], or in data that is to complex to obtain under certain clinical settings.

Convolutional Neural Networks (CNNs) are robust enough to learn deep properties of the images that are usually ignored but are relevant for the objective task, avoiding intermediate steps such as segmentation or feature design and selection [108]. However, this advantage is achieved at the cost of learning from extremely large, annotated training sets [87]. Thus, their applicability is restricted to domains

that have large amounts of labelled data, which is normally difficult to obtain as image acquisition and labelling require experts and costly devices [22]. To the best of our knowledge, CNNs were only used once for glaucoma detection, in the work by Chen *et al.* [35]. Such a method requires one to previously segment the ONH and preprocess the resulting image, and by means of training the CNN on an augmented version of a large local data set, the authors were able to automatically learn relevant features that were suitable for glaucoma identification.

As an alternative to an expensive and dedicated training task, different authors have recently proposed to use off-the-shelf CNNs to solve problems different from those for which they were originally trained [22, 39, 183, 199]. That is, CNNs that were pre-trained using a large amount of data from a different problem are transferred to extract high-dimensional but discriminative feature vectors from images from a completely different domain. These features are afterwards used to train a dedicated classifier designed to solve the new task. By means of such an approach it is possible to solve different computer vision problems using the same, single model, learned from a sufficiently large, off-the-shelf training set [184]. As the CNN has already learned discriminative features, the dedicated classifier can be trained from a smaller, application-related set. Despite the fact that this strategy looks promising and was already explored for several specific computer vision applications [183], only few studies describe the application of this concept for medical image classification [22, 39, 184, 199], and none for fundus image analysis.

In this chapter we analyze the viability of using features extracted with pre-trained, off-the-shelf CNNs in the context of glaucoma detection in limited-size fundus image data sets. Different levels of cropping around the ONH area combined with state-of-the-art preprocessing strategies, such as vessel inpainting and contrast enhancement, are explored as inputs to a CNN in order to evaluate its capability to better characterize the disease. Thus, instead of using manually engineered features, we propose to use only simple preprocessing methods to transfer the already trained CNN to this other task. ℓ_1 and ℓ_2 regularized logistic regression classifiers are trained with these features to obtain a likelihood indicating the presence or the absence of glaucoma in each image. Results on a publicly available set of fundus images suggest that this approach yields significant discrimination between glaucomatous and

healthy eyes.

4.2 Methods

We explore how to transfer pre-trained CNN from non-medical data to the task of glaucoma detection without relying on an expensive retraining using large amounts of data. In general, this problem is known as transfer learning [224], and can be tackled by following one of these two different approaches. One consists of feeding the pre-trained network with the new images, retrieving the outputs of the first fully connected layer as feature vectors [183], and using them to train a new classifier explicitly devoted to the new task. The second strategy is to not only replace and retrain the classifier on top of the CNN on the new data set, but to also fine-tune the weights of the pretrained network by continuing the backpropagation process. Depending on the new task, it is possible to fine-tune all the layers of the net or to keep some of the earlier ones fixed, restricting the fine-tuning to some of the higher-level layers of the network [184]. This is motivated by the fact that the earlier features of a CNN contain generic features that might be useful to many tasks (such as edge, color or texture detectors), but later layers become progressively more specific to the details of the classes contained in the original data set. However, fine-tuning requires a sufficiently large sample of data to avoid overfitting, a setting that cannot be accomplished with current publicly available sets for glaucoma detection.

The alternative approach of using the output features of the pre-trained network (known as *CNN codes*) and train an offline classifier based on them usually suffers if the images significantly differ from the original ones. In order to reduce the effect of this issue, we propose to analyze the usage of different preprocessing strategies before feeding the network with the images. If the preprocessing strategy enhances image characteristics that are related to the presence of glaucoma, it is expected that this approach can contribute to the computation of better features. A schematic diagram of our approach is illustrated in Figure 4.1. We describe both the preprocessing methods and the classifier in the sequel.

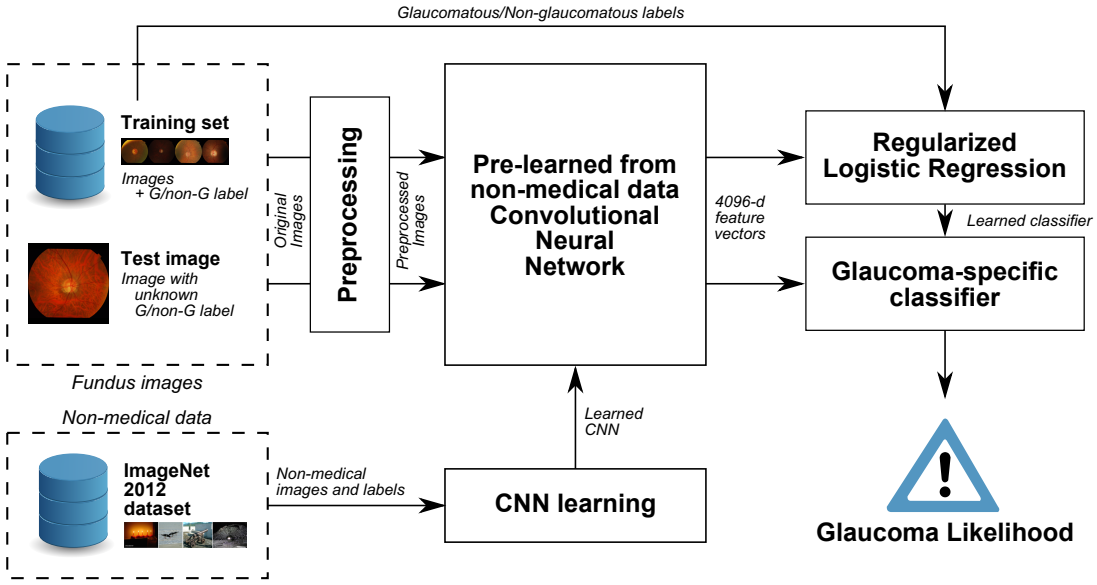


Figure 4.1: Schematic representation of our method for transferring CNNs pre-trained from non-medical data to glaucoma detection in fundus images.

4.2.1 Image preprocessing

Fundus images are preprocessed with state of the art techniques before extracting CNN codes. Figure 4.2 shows examples of all the operations we consider. In particular, we evaluate the contribution of cropping around the area of the ONH, applying contrast enhancement (CLAHE), and including or removing the vessels.

4.2.1.1 Zooming around the ONH

Glaucoma manifests specifically on the ONH region [27]. As images are usually downscaled before feeding a CNN, it can be hypothesized that the network might not be able to capture enough information about the existence of the disease due to the reduced size of the ONH. Or, on the contrary, it might be possible that the CNN can better characterize glaucoma by capturing information from other non-ONH structures of the retina, such as the vascular tree. Several experiments considering different zooms of the original fundus images, centered around the ONH, were carried out to validate one of these hypotheses. 4 different zooms were considered (Figure 4.2): the original image as it is, including the blank regions outside the field

Figure 4.2: Preprocessing strategies evaluated. First group: without CLAHE. Second group: with CLAHE. From left to right: original image, cropped FOV, peripapillary area (PPA), and ONH. First row: original images. Second row: images after vessel inpainting.

of view (FOV); a squared crop of the area inside the FOV; the peripapillary area (PPA), which involves the ONH and its surrounding zone; and the ONH itself.

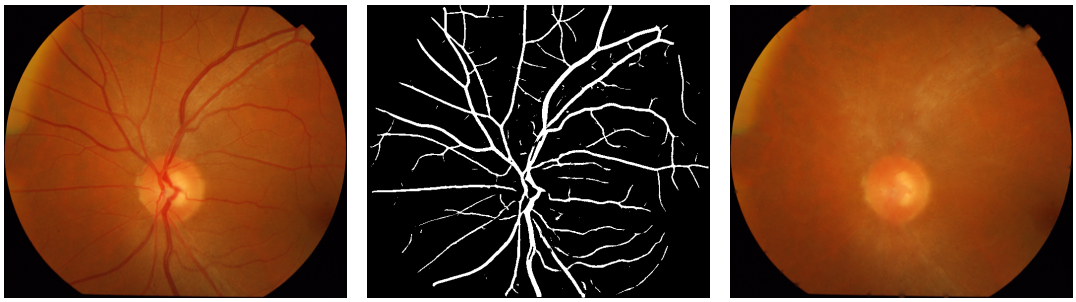
4.2.1.2 CLAHE enhancement

Contrast-Limited Adaptive Histogram Equalization (CLAHE) is an enhancement operation that is widely applied in retinal image analysis due to its capability to improve the contrast and illumination of the fundus [20] (Figure 4.2). Additionally, it is known to benefit retinal nerve fiber characterization [100, 147]. Performances with and without applying CLAHE were evaluated.

4.2.1.3 Vessel subtraction

Though there are publications indicating a certain correlation between glaucoma and changes in the vascular structures (see Section 2.4.2.2), other authors suggest to remove them before extracting features [27, 75, 124]. Based on such assumptions, we obtain features from images with and without the vessels inpainted.

The vascular tree is segmented using a slightly modified version of the method we presented in Chapter 3. In particular, pixel level features are extracted from the green band of the RGB images after CLAHE enhancement, rather than from the green band as it is. We also change the input features, using responses to line detectors [138] and to B-COSFIRE filters [20] as unary features, and Zana and Klein vessel enhanced images [226] as pairwise features. We observed that this combination



(a) DrishtiGS_041 color image. (b) Vessel segmentation (c) Image after vessel inpainting

Figure 4.3: Vessel subtraction.

of features empirically achieved better binary representations of the vascular trees in our data sets.

Vessels are removed from the original color image by applying inward interpolation on each of the color bands: new intensities are smoothly interpolated inward from the pixels values on the outer boundary of the vascular binary mask. As this strategy is affected if the segmentation method underestimated the actual vessel calibre (because the iterative inpainting might reintroduce the vasculature to the image instead of removing it), the vascular masks are morphologically dilated using disks of size 3 pixels. Figure 4.3 illustrates a segmentation of the vascular tree and the resulting image after removing it.

4.2.2 ℓ_1 and ℓ_2 regularized logistic regression

Glaucoma detection is a binary classification task that we estimate by means of a supervised learning approach known as regularized logistic regression [77]. Let S be a training set composed of n training instances $\{(\mathbf{x}_i, y_i), i = 1, ..., n\}$. Each $\mathbf{x}_i \in \mathbb{R}^d$ is a d-dimensional feature vector extracted from the i-th image using a CNN, and $y_i \in \{-1, +1\}$ represents its corresponding class label, $+1$ indicating glaucomatous and -1 non-glaucomatous. Logistic regression can be written as a regularized risk minimization with logistic loss. The objective function to be minimized is:

$$\hat{\beta} = \arg\min_{\beta} \lambda\Omega(\beta) + \sum_{i=1}^{n} \log\left(1 + e^{-y_i\langle\beta,\mathbf{x}_i\rangle}\right) \tag{4.1}$$

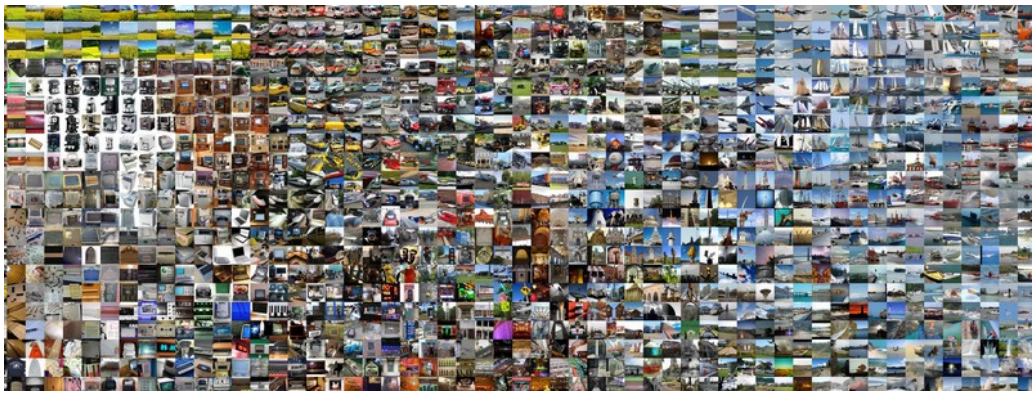


Figure 4.4: A sample of the natural images from ImageNet 2012 used for training both OverFeat and VGG-S.

where β is a coefficient vector of the linear discriminant function, $\lambda \geq 0$ is a scalar parameter controlling the degree of regularization by the regularizer $\Omega : \mathbb{R}^d \mapsto \mathbb{R}_+$ and $\langle \cdot, \cdot \rangle$ is the canonical inner product in $\mathbb{R}^d$. Two well-known regularizers were used as Ω, the $\ell_1 = \| \cdot \|_1$ and $\ell_2 = \| \cdot \|_2$ norms. ℓ_1 imposes the sum of the absolute values of the parameters β to be small, encouraging sparse parameter vectors [135]. This setting makes this regularizer suitable for simultaneous learning and feature selection when the size of the training set n is small and the size of the feature vector d is high, as in this case. By contrast, ℓ_2 regularization may not be effective when only a few features are relevant [135]. Both norms were experimentally validated to evaluate if the high dimensional CNN codes are discriminative by themselves, or if only some of their components are relevant for glaucoma identification. The likelihood of a given image having glaucoma or not is obtained by the linear discriminant function $\hat{y} = \langle \beta, \mathbf{x}_i \rangle$. An existing implementation of k-support regularized logistic regression [18, 185] was used to minimize the objective function, assigning $k = 1$ for the ℓ_1 norm and $k = d$ for the ℓ_2 norm.

4.3 Materials

4.3.1 Convolutional neural networks

Two different CNNs were used in our experiments, namely OverFeat [181][1] and VGG-S [33][2]. Both networks were initially trained on the *ImageNet 2012* training set (Figure 4.4), which comprises 1.2 million natural, non-medical color images labeled with 1000 classes, with the second one reporting the highest performance in the PASCAL-VOC competition.

OverFeat architecture is inspired by the work of Krizhevsky *et al.* [103], with minor changes described by Sermanet *et al.* [181]. We use the accurate version of their feature extractor, which is composed of 6 convolutional layers, with filters sizes from 7×7 to 3×3 containing from 96 to 1024 kernels. Techniques such as max-pooling, rectification non-linearities (ReLUs) and dropout are used, as described in the original reference. A similar architecture was introduced by Chatfield *et al.* [33], known as VGG-S. It comprises 5 convolutional layers instead of 6, with filters sizes similar to those in OverFeat but with 512 kernels in the last 3 convolutional layers. The last three layers of both CNNs are fully connected, and they are responsible of scoring the likelihood of each of the classes. As our purpose is to use these CNNs as feature extractors, the output of the first fully-connected layer is used here to retrieve 4096-d feature vectors from input images. Images have to be rescaled to 231×231 or 224×224 pixels before feeding Overfeat and VGG-S, respectively.

4.3.2 Data sets

Experiments were carried out on the data set DRISHTI-GS1 [187]. It consists of 101 fundus images, taken with eyes dilated, centered on the ONH, at 30° FOV, with 2896×1944 pixels and saved as uncompressed PNG images. All images were originally collected at the Aravind Eye Hospital, India. Glaucoma patient selection was done by clinical investigators based on clinical findings during examination. Selected patients were 40-80 years of age, with a roughly equal number of males and females. Patients undergoing a routine refraction test and not found to be

[1] `http://cilvr.nyu.edu/doku.php?id=software:overfeat:start`
[2] `http://www.vlfeat.org/matconvnet/pretrained/`

glaucomatous were chosen to represent the normal class. The diagnostic opinion for each image (normal or glaucomatous) was obtained from 4 glaucoma experts based only on the images, and the provided gold standard was derived as the majority opinion, i.e. 3 out of the 4 experts.

For vessel inpainting, vascular trees were segmented with a model trained on the DRIVE [141] training set and calibrated using the DRIVE test set. After model selection, parameters were set to $C = 10^2/c$, with c equal to the number of training pixels, and $\theta_p = 5$. Feature parameters were fixed to the values provided in the original references. As features were originally designed for DRIVE and depend on the resolution of the images–which significantly differs between data sets–DRISHTI-GS1 images were downscaled by a factor of 2.96 before segmentation so vessels appear at a similar resolution to DRIVE. As images are then further downsampled before applying CNNs, this process does not negatively affect the end result.

As all these data sets do not include FOV masks, which are necessary for further processing, we automatically generate them by thresholding the luminosity plane of the CIELab version of the RGB images at a value of 0.01[20]. If the resulting binary mask is all ones, an alternative approach is applied to estimate the FOV mask, were the RGB bands are summed up and the resulting image is thresholded at 150 level. To smooth borders and reduce noise, the mask is postprocessed with a median filter using square windows of side 5 pixels, and only its largest connected component is preserved.

4.4 Results

All the preprocessing strategies were analyzed in combination with the OverFeat network with the purpose of determining which of them is the best. Due to the relatively small size of the available data set, each configuration is evaluated with 200 trials as follows. First, images are randomly divided into training (70%) and test sets (30%), as suggested in the literature [77]. 30% of the training images are randomly selected and separated as the validation set, which is used to determine the value of the regularization parameter $\lambda = 10^i$ of the logistic regression classifier, with $i \in \{-5, -4, ..., 6\}$. The area under the ROC curve (AUC) is used as a numeri-

cal indicator of the overall model performance during model selection. Two different techniques for augmenting the training sets are evaluated: images are rotated at two different angles (90° and 45°) until reaching 360°, and horizontally flipped, incrementing the size of the sets by a factor of 8 and 16, respectively. Performance of each configuration is measured in terms of the average area under the ROC curve [58]. Results are presented in Figure 4.5 and analyzed in the sequel.

If vessels are not inpainted, OverFeat codes achieve the best results with images cropped in the FOV region, independently of the regularizer used. By contrast, if vessels are removed from the images, then both norms report the best performance when images are zoomed in the PPA. In any of these cases it is not possible to determine which is the best augmentation technique, as it varies from one configuration to another. In the case of non-inpainted cropped FOV images, the best AUC (0.7626) is obtained by OverFeat features with images augmented using 90° rotations; when using inpainting, the largest value (0.7212) is achieved without augmentation. CLAHE enhancement is ill advised in most of the analyzed configurations, as it significantly decreases the AUC values. It only demonstrates a positive contribution under vessel inpainting, in the cropped FOVs and the ONHs, and when using the ℓ_1 norm with features obtained from the original images. The ℓ_2 norm performs equally or better than the ℓ_1 for almost all the settings. With regards to vessels, their subtraction does not offer any improvement in results. In fact, most of the AUC values are decreased when this operation is used. The only exception can be observed when focusing on the ONH and applying the CLAHE operation. An additional experiment, not included in the figure, was made feeding the classifiers with OverFeat codes extracted from the raw binary segmentations of the vessels, but AUCs obtained did not significantly differ from chance performance (the highest was 0.5303 in the 90° augmented version).

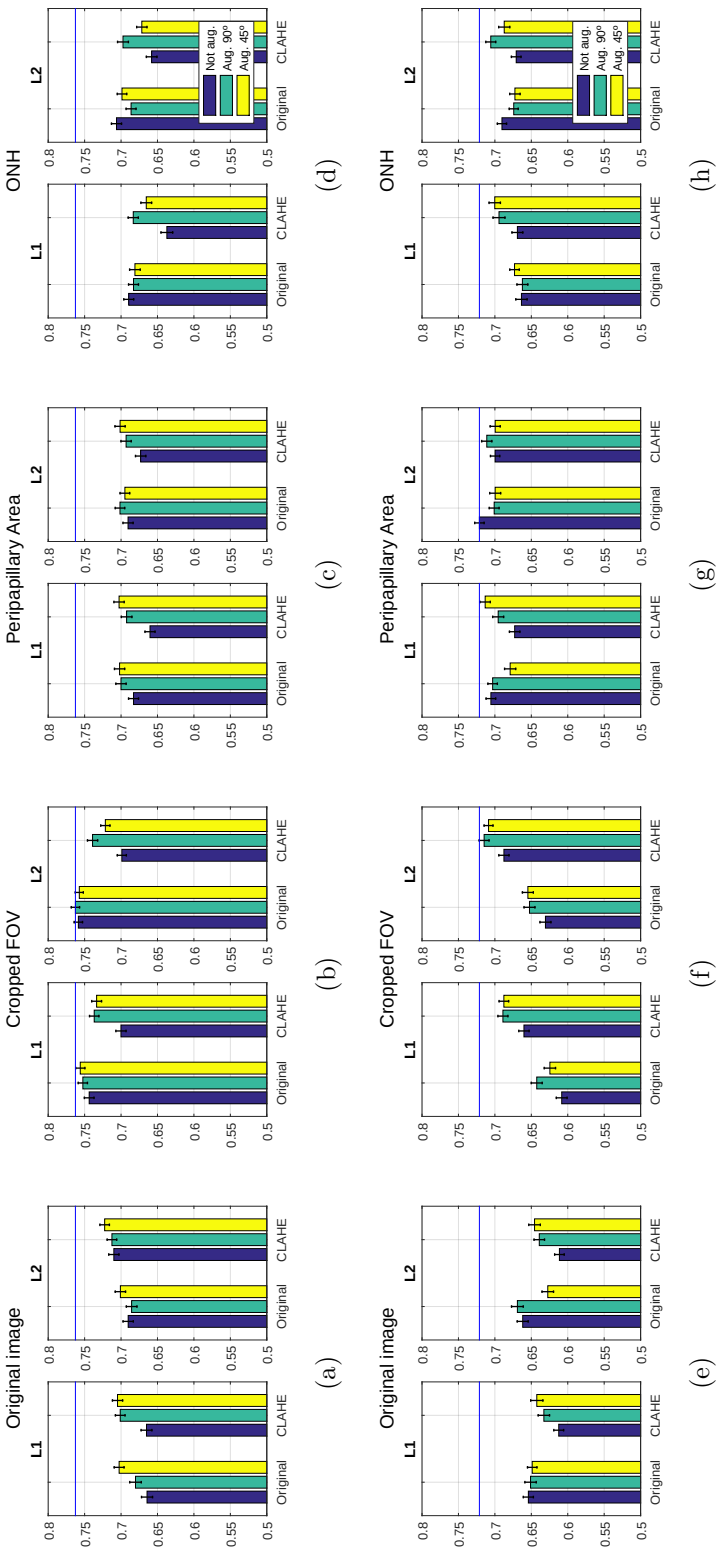


Figure 4.5: Area under the average ROC curves obtained from 200 trials [58] using OverFeat features with ℓ_1 or ℓ_2 regularized logistic regression, respectively. First row: using images without vessel inpainting. Second row: using images with vessel inpainting. The blue line corresponds to the best AUC value for the row.

An additional experiment was performed using VGG-S, feeding it with images preprocessed with the best strategy we found when using OverFeat (RGB images cropped inside the FOV without CLAHE, without inpainting, 90° augmentation; and RGB images cropped around the peripapillary area without CLAHE, with vessels inpainted and without data augmentation). The purpose of this analysis was to determine if slight changes in the architecture of the CNN might affect the performance of the transferred network. In both experiments we used ℓ_2 regularized logistic regression due to its better results while evaluating preprocessing techniques. We observed that OverFeat features performed better than VGG-S both in images with (AUC= 0.7212 vs. AUC= 0.6655, respectively) and without vessels inpainted (AUC= 0.7626 vs. AUC= 0.7180, respectively).

Extensive comparison with respect to other state of the art strategies is not feasible as most of them used their own private data sets. To the best of our knowledge, only the work by Chakrabarty *et al.* [32] was evaluated on DRISHTI-GS1, reporting an AUC value of 0.78. Such an approach is based on a classification framework that fuses both overall image and segmentation based features within a co-training based semi-supervised setting. The main difference with our work is that we only rely on features from the cropped version of the original fundus image, and we do not integrate any information extracted from ONH segmentations. Additionally, a larger training set was used for learning the classification model reported in that work.

4.5 Discussion

In this chapter we have analyzed the capability of transferring pre-trained, off-the-shelf CNNs for glaucoma detection using fundus photographs. A variety of state-of-the-art preprocessing techniques, including different crops around the ONH region, were used to feed the neural networks and to recover features that are then used in two different regularized logistic regression models. Performance on cropped FOVs was higher than in other crops when using OverFeat, indicating that such a network might be able to recover valuable information from other regions than the ONH. This evidence is also supported by the low AUC values achieved when vessels are inpainted. These results contradict other studies that focused on the extraction of

features only in the PPA or in the ONH after removing the vessels [26, 27]. The negligible AUC values obtained using only the vasculature, however, suggest that vessels are not useful by themselves but only when combined with other sources of information in the fundus. CLAHE enhancement, which is known to improve the contrast of the fundus and is widely applied, increases the AUC values under some settings. However, in most of the non-inpainted cases, it is not helpful. By contrast, when vessels are removed, CLAHE benefits results achieved on cropped FOVs and in the ONH, and also in the PPA when ℓ_1 regularization is used. This might be associated with an improvement in the discrimination of the optic cup and the ONH, though such an enhancement is not competitive with results achieved using other settings. ℓ_1 regularization is known as a suitable tool for simultaneous feature selection and classification, as it favors sparse coefficient vectors, a feature that is not provided by ℓ_2 regularization, which tends to use all the coefficients. As the ℓ_2 norm performs equal to or better than the ℓ_1 norm in most of the experiments, it can be conjectured that all the coefficients in the feature vectors are useful for this task. However, it worth mentioning that the improvement obtained combining CLAHE with the ℓ_1 norm can be associated with an improvement of certain coefficients in the feature vector. Additionally, we observed that in this data set VGG-S features do not perform better than OverFeat ones.

In conclusion the preprocessing strategy, although being simple, supports transfer learning from small data sets where fine-tuning via retraining is not possible. In terms of the overall performance of our method, we achieved a competitive area under the ROC curve score with respect to other existing strategies [32], but with the advantage of not requiring the segmentation of the optic cup and the optic disc regions, and training from a smaller data set. Nevertheless, it is important to underline that further experiments on larger data sets must be done to validate the generalization ability of our method to other image characteristics.

Chapter 5

Red lesion detection for DR screening

As previously stated in Chapter 2, the earliest signs of diabetic retinopathy are red lesions, a general term that groups both microaneurysms (MAs) and hemorrhages (HEs). In daily clinical practice, these lesions are manually detected by physicians using fundus photographs. However, this task is tedious and time consuming, and requires an intensive effort due to the small size of the lesions and their lack of contrast. Computer-assisted diagnosis of DR based on red lesion detection is being actively explored due to its improvement effects both in clinicians consistency and accuracy. Moreover, it provides comprehensive feedback that is easy to assess by the physicians. Several methods for detecting red lesions have been proposed in the literature, most of them based on characterizing lesion candidates using hand crafted features, and classifying them into true or false positive detections. Deep learning based approaches, by contrast, are scarce in this domain due to the high expense of annotating the lesions manually.

In this chapter we propose a novel method for red lesion detection based on combining both deep learning and domain knowledge. Features learned by a convolutional neural network (CNN) are augmented by incorporating hand crafted features. Such ensemble vector of descriptors is used afterwards to identify true lesion candidates using a Random Forest classifier. We empirically observed that combining both sources of information significantly improve results with respect to using each

approach separately. Furthermore, our method reported the highest performance on a per-lesion basis on DIARETDB1 [97] and e-ophtha [48], and for screening and need for referral on MESSIDOR [49] compared to a second human expert. An extensive analysis of the complementarity of the deep learned features with respect to the hand crafted ones is also provided, with the purpose of assessing their contribution in the discrimination process. Results highlight the fact that integrating manually engineered approaches with deep learned features is relevant to improve results when the networks are trained from lesion-level annotated data.

This chapter is organized as follows. Section 5.1 provides insights about the problem and current literature in the area. In Section 5.2 we describe our method in detail, explaining our candidate detection strategy, the CNN architecture, the features used to complement it and the Random Forest classifier. Section 5.3 describes our experimental setup, and Section 5.4 presents the results obtained. Section 5.5 analyzes those results in details and the complementarity of both sets of features. Finally, Section 5.5 concludes the chapter.

The implementation of our method is made publicly available in `https://github.com/ignaciorlando/red-lesion-detection`. The work presented in this chapter is described in our preprint:

- J. I. Orlando, E. Prokofyeva, M. del Fresno, and M. B. Blaschko. Learning to detect red lesions in fundus photographs: An ensemble approach based on deep learning. *arXiv preprint arXiv:1706.03008*, 2017.

5.1 Motivation

Although fundus photographs are currently the most economical non-invasive imaging technique for DR screening, manual diagnosis requires an intensive effort to screen the images [131]. Red lesions appear as small red dots that might be subtle and too small to be detected at first glance (Figure 5.1). Large HEs, on the contrary, are more evident and less difficult to visualize.

Automated methods for computer-aided diagnosis are known to significantly reduce the time, cost, and effort of DR screening: their high throughput ensures the more efficient analysis of large populations [176]. They also reduce the intra-

Figure 5.1: Examples of red lesions observed in fundus photographs from DIARETDB1 [97].

expert variability, which is commonly high due to the small size and the irregular shape of the lesions [3]. These systems are usually aided by an automated module for red lesion detection. In general, the problem of red lesion detection is tackled using a two-stage approach, consisting first of detecting a set of potential candidates, and then refining this set with a classifier trained using hand crafted features [142, 143, 180, 207].

Convolutional Neural Networks (CNNs) have recently emerged as a powerful framework to solve a large variety of computer vision and medical image analysis problems [103, 202, 230]. Such methods are able to learn features automatically from a sufficiently large training set, without requiring the manual design of the filters. CNNs are known to outperform other manually engineered approaches on a large variety of applications [169]. Their discrimination ability is usually affected by the amount of available training data: deeper architectures are known to be able to learn more discriminative features, although at the cost of requiring larger data sets to prevent overfitting and ensure a proper generalization error [71]. Image level annotations of large scale data sets can be obtained in a relatively economical way [196]. However, labeling images at a lesion level is costly, tedious and time consuming, as it requires the intervention of experienced experts who must zoom within different areas of the images to identify every single pathological structure, as accurately as possible. This fact significantly influences the performance of deep learning based approaches for red lesion detection, which must be trained using lesion level annotated data.

In this study we propose to take advantage of both deep learned and manual engineered features for red lesion detection in fundus photographs. In particular, we propose to learn a first set of discriminative features using a light CNN architec-

ture, and then augment their original characterization ability by incorporating hand crafted descriptors. These ensemble vectors of features are used to train a Random Forest classifier that is applied at test time to discriminate between true and false lesion candidates. We experimentally observed that the deep learned features are complementary to the manually engineered, and are aided by the incorporation of domain knowledge.

Red lesion detection in fundus photographs have been extensively explored in the literature, although most of the existing approaches are based on detecting MAs or HEs separately, and not both structures simultaneously [142, 180, 200]. Moreover, current existing approaches are based exclusively on hand crafted features. This is likely due to the fact that deep learning based methods have to be trained from large data sets with lesion level annotations. This setting has direct implications on why deep learning based models have been ignored for tackling the problem of red lesion detection. One exception is the method for HEs detection by van Grinsven *et al.* [200]. This approach is focused on detecting HEs at different scales, which are in general more evident than MAs. By contrast, our method is used for detecting both MAs and small HEs simultaneously, which are more difficult to be visually assessed by physicians.

In this study we present an ensemble approach that improves the features learned by a CNN by incorporating domain knowledge. Only few efforts have been made in the literature to analyze the viability of such an approach. Annunziata *et al.* [15], for instance, propose to initialize a convolutional sparse coding approach with manually designed filters to accelerate its learning process and improve their original discriminative power. That approach is applied for detecting curvilinear structures such as neurons or retinal vessels, which are easier to manually trace. Venkataramani *et al.* [202] have observed that state of the art descriptors significantly improve the performance of transferred CNN features when applied to kidney detection in ultrasound images. The main difference with respect to our approach is that our CNN is trained from scratch from a domain specific data set, while the approach presented in [202] is based on fine-tunning a CNN trained from natural images.

From our literature review, we identified two main methods resembling our approach, although with different applications and based on different CNN architec-

tures. Zhen *et al.* [230] introduce a method for identifying landmarks in 3D CT scans using the output of a dedicated CNN in combination with Haar features to boost the quality of the results. Its deep learning based component is divided into two stages: a first stage, based on a light architecture with only one hidden layer, is used to recover a large set of landmark candidates; the second stage, made up of three hidden layers and trained using sparsity priors, is used to recover a large vector of neural network features, which is combined with Haar features to train a probabilistic boosting-tree classifier. In order to save as much data as possible for training the CNN and the lesion classifier, we avoided performing candidate detection in a supervised way. Instead, a combination of morphological operations and image processing techniques is used to retrieve potential lesions, without using training data. This allows us to train a slightly deeper architecture in the subsequent stage, only dedicated to classifying the lesion candidates, which is able to capture discriminative features from the training patches.

The method for mitosis detection on histopathology images presented in [209] is also similar to ours. It uses candidate detection as well, and a RF classifier trained using hand crafted features is applied to assign a probability of being a true mitosis candidate. In parallel, a CNN with two convolutional layers and one fully connected layer is trained from patches around the candidates to retrieve an additional probability. The final decision is performed via consensus of the predictions of the two classifiers by weighting both probabilities using two manually tuned parameters. We took the alternative approach of using both feature vectors simultaneously to train the RF classifier, as it can take advantage of the interaction between both the deep learned and the hand crafted features.

5.2 Methods

In this chapter we propose to learn discriminative models for red lesion detection by combining both deep learned and hand crafted features. A schematic of our method is depicted in Figure 5.2. First, an unsupervised, candidate detection approach based on morphological operations is applied to retrieve a set of potential lesions (Section 5.2.1). Next, a CNN is trained from a set of regular patches centered on

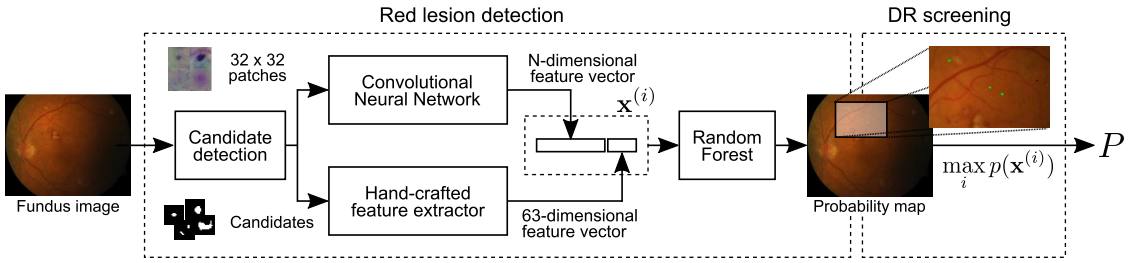


Figure 5.2: Overview of our method for red lesion detection.

each candidate connected component to learn a first feature vector (Section 5.2.2). These descriptors are augmented with a set of 63 hand crafted features to improve their ability to distinguish the true positive lesions (Section 5.2.3). A Random Forest (RF) classifier is trained using this hybrid feature vector, and is then applied for refining the set of candidates, discriminating between true lesions and false positives (Section 5.2.4). Since the presence of red lesions is the first indicator of DR, the maximum over lesion likelihoods is used to assign a DR probability, as done in [180] and [16].

5.2.1 Candidate detection

Our strategy for candidate detection is illustrated in Figure 5.3. First, the green band G from the original color image I is taken, since it is the one that allows a better visual discrimination of the red lesions. To avoid artifacts in the borders of the field of view (FOV) that might hide potential lesions (Figure 5.4(b)), a wider aperture of $\frac{3}{30}\mathcal{X}$ pixels is simulated [188] from G, where $\mathcal{X}$ corresponds to the width in pixels of the FOV. Since our purpose is to develop a system sufficiently general to be applied at different image resolutions, all the relevant parameters are expressed in terms of $\mathcal{X}$.

As uneven background illumination might hide potential lesions occurring within the darkest areas of the images, a r-polynomial transformation is applied on pixel intensities:

$$I_{\mathrm{W}}(i,j) = \begin{cases} \frac{\frac{1}{2}(u_{\max}-u_{\min})}{(\mu_W(i,j)-\min(G))^r}, & G(i,j) \leq \mu_W(i,j) \\ \\ \frac{-\frac{1}{2}(u_{\max}-u_{\min})}{(\mu_W(i,j)-\max(G))^r}, & G(i,j) > \mu_W(i,j) \end{cases} \tag{5.1}$$

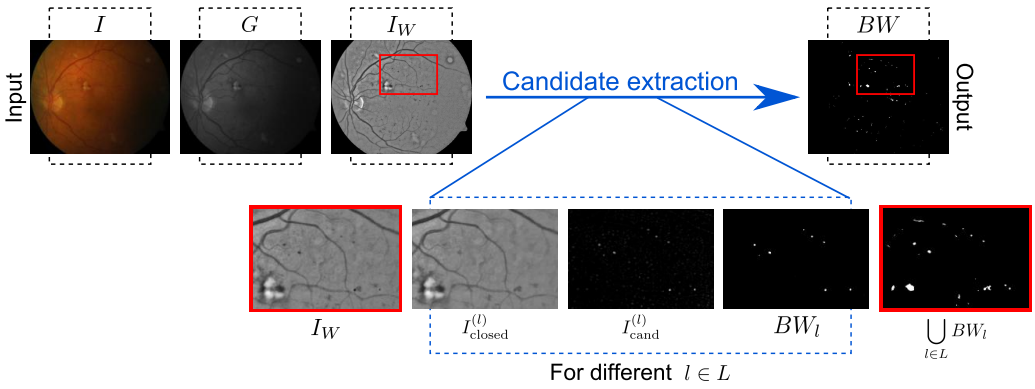
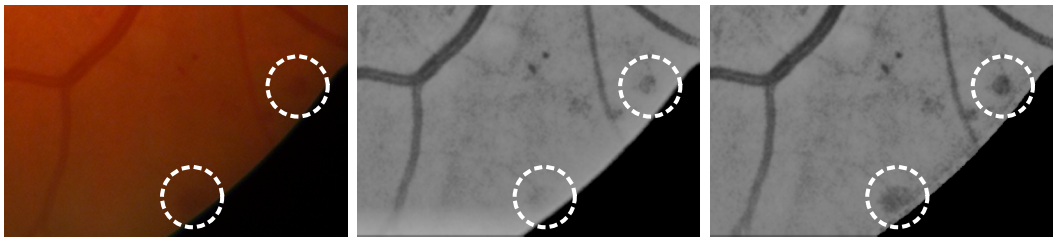


Figure 5.3: Red lesion candidate detection. See Section 5.2.1 for a detailed description of the process.



(a) Original RGB image (b) I_W from G (c) I_W from expanded G

Figure 5.4: Effect of the FOV expansion on the lesion candidates located closely to the border of the FOV.

with $r = 2$, $u_{\min} = 0$ and $u_{\max} = 1$, respectively, and where μ_W is the local average intensity on square neighborhoods of length W, computed for each (i, j) pixel [207]. We observed that using $W = 25$ performed sufficiently well for enhancing images with 536 pixels of horizontal resolution such as those in the DRIVE data set [141], so this parameter is automatically adjusted using $W = \frac{25}{536}\mathcal{X}$. Figure 5.4 illustrates how the expansion of the FOV border and the subsequent intensity transformation improve the contrast of subtle lesions located in the border of the FOV.

A Gaussian filter with $\sigma = \frac{5}{536}\mathcal{X}$ is applied to I_{W} to reduce noise, resulting in a new image I'_{W}. Afterwards, different morphological closings are performed on I'_{W} using linear structuring elements of length $l \in L$ at angles θ spanning from 0 to 180° with increments of 15°. The set of relevant scales L is a fixed parameter that is also automatically adjusted in terms of $\mathcal{X}$, as explained in Section 5.3.2. By taking

the minimum response over all the considered angles, an image $I^{(l)}_{\text{closed}}$ is obtained in which responses to lesions with sizes smaller than l were reduced, and all the remaining structures are still preserved [207]. A score map is then obtained by:

$$I^{(l)}_{\text{cand}} = I^{(l)}_{\text{closed}} - I_{\text{W}}. \tag{5.2}$$

Afterwards, a thresholding operation is applied on $I^{(l)}_{\text{cand}}$, where the threshold is automatically determined in such a way that a maximum of $K = 120$ candidates are retrieved from the score map. In order to achieve this goal, thresholds t_s from $\min(I^{(l)}_{\text{cand}})$ to $\max(I^{(l)}_{\text{cand}})$ with increments of 0.002 are explored until the number of connected components in the resulting binary maps is less than or equal to K. To support the cases in which no lesions are detected or when is not possible to detect less than K candidates, a lower bound t_l and an upper bound t_u are experimentally set such that:

$$t_K = \begin{cases} t_l, & \forall t_s : \text{CC}(I^{(l)}_{\text{cand}} > t_s) < K \\ t_k, & \text{CC}(I^{(l)}_{\text{cand}} > t_s) \leq K \\ t_u, & \forall t_s : \text{CC}(I^{(l)}_{\text{cand}} > t_s) > K \end{cases} \tag{5.3}$$

where CC is a function than counts the number of connected components in the thresholded score map. Once t_K is fixed, a binary map of candidates is obtained by thresholding $BW_l = I^{(l)}_{\text{cand}} > t_K$ [207]. This operation is repeated for different values of $l \in L$ to capture potential lesions at different scales, so the binary map of candidates BW is obtained as $BW = \bigcup_{l \in L} BW_l$. Finally, as BW might include small candidates which usually are not associated to any pathological region but with noise, all connected structures in BW with less than px pixels are discarded. The automated model selection procedure used to set the values of K and px and the scales in L is described in Section 5.3.2.

Figure 5.5 presents a random sample of the potential lesions retrieved by the method on a randomly selected image from DIARETDB1 training set. It is possible to see that most of false positive samples correspond to vascular branching or crossing points, vessel segments and beadings, scars due to laser photocoagulation or black spots of dirt in the capture device, as reported in [180]. This setting underlines the importance of refine the candidates to remove false positives.

5.2.2 CNN-based features

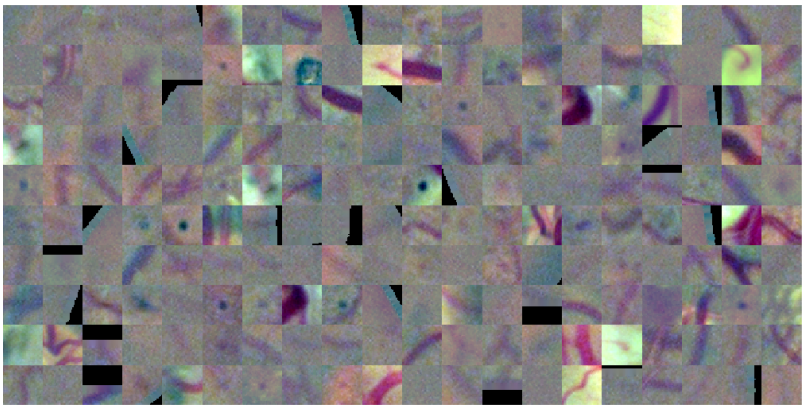
We train a dedicated CNN to characterize each red lesion candidate. For this purpose, each color band of the original image I is equalized first as proposed in [200]:

$$I_{\text{ce}}(i,j;\sigma) = \alpha \cdot I(i,j) + \tau \cdot \text{Gaussian}(i,j;\sigma) * I(i,j) + \gamma \tag{5.4}$$

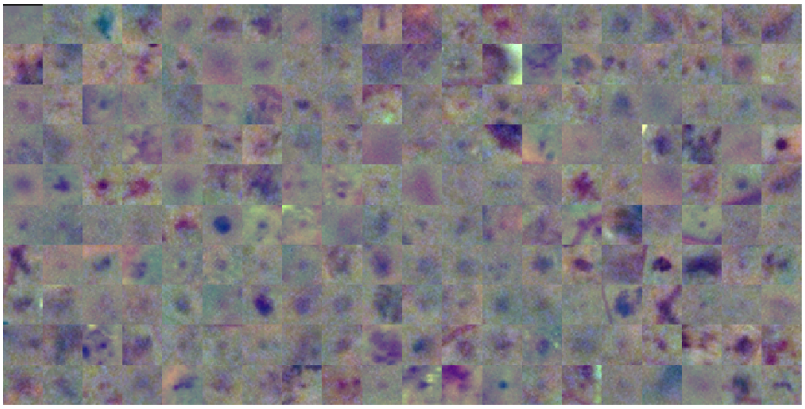
where $*$ is a convolution, the Gaussian filter has a standard deviation $\sigma = \frac{\mathcal{X}}{30}$, and $\alpha = 4$, $\tau = -4$ and $\gamma = 128$ were set following [200]. We empirically observed that this preprocessing operation not only dramatically diminishes the number of epochs needed for training but also improves the discrimination ability of the CNN. As explained in Section 5.2.1, a wider FOV is also simulated for each color band to prevent any undesired effect in the FOV border.

A training set $S = \{(X^{(i)}, y^{(i)})\}, i = 1, ..., n$ is built for training the CNN, where each sample $X^{(i)}$ is a patch around the center of each red lesion candidate, as extracted from I_{ce} (Figure 5.5). The patch size is taken as double the length of the major axis of the candidate to recover not only the candidate itself but also its surrounding area. This allows the CNN to capture both internal features of the candidate and also information about its shape, borders and context. To reduce the number of parameters of the network, all the patches are normalized to regular windows of 32×32 pixels. Samples are centered by the training set mean. The label $y^{(i)} \in \{0, 1\}$ associated to the candidate is assigned according to the ground truth labeling on the data set: if the candidate overlaps with a true labeled region, then the window is assumed to be a true red lesion ($y = 1$); if it does not, then it is assumed to be a false positive ($y = 0$).

The CNN is trained from scratch on an 8x augmented version of this training set, obtained by flipping images horizontally and rotating them by $90°, 180°$ and $270°$. Its architecture is depicted in Table 5.1. It comprises 4 convolutional layers and 1 fully connected layer with 128 units. This layer is used to retrieve the 128-dimensional vector of deep learned features. The CNN was designed by using the original LeNet architecture as the initial baseline, and introducing changes by evaluating their contribution on reducing the empirical error on a held-out validation set, which was randomly sampled from the training set. We also incorporated dropout after the first convolutional layer. Using deeper architectures did not empirically



(a) Non-lesions (false positive candidates)



(b) Lesions (true positive candidates)

Figure 5.5: CNN training set. Random sample of 200 patches for (a) non lesions and (b) true lesions. See Section 4.3.1 for details of the construction of the training set.

improve results.

Depending on the number of images in the training set with an advanced level of DR and the size of the lesions we focus on detecting, the classification problem is imbalanced to a greater or lesser degree. Hence, the proportion of true positive lesions might be significantly smaller that the number of false positive ones. If this imbalance grows dramatically, it was previously observed that the typical cross-entropy loss is affected, and, as a consequence, fewer true positives are retrieved [119]. Thus, we

used a class balanced cross-entropy loss, given by:

$$\mathcal{L}_\beta(\mathbf{W}) = -\beta \sum_{i \in Y_+} \log P(y^{(i)}|X^{(i)}; \mathbf{W}) - (1-\beta) \sum_{i \in Y_-} \log P(y^{(i)}|X^{(i)}; \mathbf{W}))$$

where $\mathbf{W}$ are the weight parameters of the network; P is the probability obtained by applying a sigmoid function to the activation of the fully connected layer; Y_+ and Y_- are the subsets of true and false positive samples, respectively; and $\beta = |Y_-|/(|Y_+| + |Y_-|)$ is the ratio of negative vs. positive samples in S. This cost function was optimized using stochastic gradient descent, with a batch size of 100 samples.

Standard dropout after all the convolutional layers was analyzed as an alternative to the reported architecture, although it was observed that it did not improve results on the validation set. Moreover, using such an approach increased the training time significantly. We noticed instead that using dropout after the first convolutional layer with a high keep probability slightly improve results. We also used weight decay of 5×10^{-4} for regularization, to penalize large $\mathbf{W}$ values during backpropagation. The combination of both dropout and weight decay showed to be sufficient in our experiments. The learning rate was initially fixed in $\eta = 0.05$, and it was divided by a factor of 2 every time that the relative improvement in current $\mathcal{L}_\beta(\mathbf{W})$ value was less than 1% of the average loss over the last 10 epochs. Finally, all the weights on the neural network were randomly initialized using a Gaussian distribution with a standard deviation of 0.05, with the only exception of the first convolutional layer, for which we used a standard deviation of 0.01.

5.2.3 Hand-crafted feature extraction

As a complementary source of information with respect to the CNN features, a 63 dimensional feature vector of hand-crafted features (HCF) is extracted per each lesion candidate and incorporated to our feature vector. Some of these descriptors were extensively explored in the literature [142, 143, 180], while other are introduced here to improve the existing ones. In general, they can be divided into two categories: intensity based and shape based features (Table 5.2).

Intensity features exploit the visual properties of the candidate areas, and are

Table 5.1: CNN architecture. Convolutional layers (conv) indicate width, height and depth of each learned filter. Pooling layers (pool) include the dimension of the pooling operation and the stride. Dropout is only applied after the first convolutional layer with a low dropout probability.

Block	**Layers**	**Filter size**	**Output size**
1	conv	$5 \times 5 \times 3$	32
	maxpool	3×3 - stride $= 2$	
	dropout	$p = 0.01$	
2	conv	$5 \times 5 \times 32$	32
	avgpool	3×3 - stride $= 2$	
3	conv	$5 \times 5 \times 32$	64
	avgpool	3×3 - stride $= 2$	
4	conv	$4 \times 4 \times 64$	128
5	fully connected	128	128
6	$\mathcal{L}_\beta(\mathbf{W})$	128	2

extracted from different versions of the color image I, obtained by applying different preprocessing strategies. In particular, we extracted descriptors used in the state of the art [142, 180] but from the following derived images:

- Original red, green and blue color bands (R, G and B, respectively).
- Green band G after illumination correction (I_W, obtained as in Section 5.2.1).
- Color bands and I_W after CLAHE contrast enhancement (R_c, G_c, B_c, I_{Wc}).
- Color bands after color equalization (R_{ce}, G_{ce}, B_{ce}).
- I_{SC}, which is the difference between the green band G and an estimated background I_{BG}, obtained using a median filter with squared windows of length $\frac{25}{536}\mathcal{X}$.
- I_{match}. This image is obtained by initially computing I_{lesion}, which is a vessel free version of I_{SC}, obtained by inpainting the vasculature as in [150], also described in Section 4.2.1.3. The difference between each pixel (i, j) in I_{lesion} and its 11×11 neighborhood is assigned to $I_{\text{match}}(i, j)$.

- $I_{\text{cand}} = \max_l I^{(l)}_{\text{cand}}$, which is the maximum response to the candidate score map described in Section 5.2.1, taken from I_W, but restricting the size of the structuring elements to the lengths $l \in \{5, 7, ..., 15\}$.

Shape based features have the ability to characterize the structure of the candidates. Red lesions are expected to be relatively circular, with small area and perimeter, and approximately equal minor and major axis. Such statistics, including compactness, eccentricity and aspect ratio, are also included as part of the domain knowledge feature vector.

We also analyzed the viability of using the segmentation of the retinal vasculature as a potential source of information. As seen in Figure 5.5(a), most of the false positive detections are located in vessel crossings or beadings. Thus, we compute an initial vessel segmentation using the method reported in Chapter 3, and postprocessing the output by removing every spurious connected component with less than $\frac{100}{536}\mathcal{X}$ pixels [150]. A morphological closing with a disk of radius 2 is afterwards applied to fill any gap due to the central reflex in arteries. Then, we measure the ratio of pixels in the candidate region that overlap with the segmentation, divided by the number of pixels in the candidate. Figure 5.6 illustrates the process of computing this feature. It can be seen that most of the false positive lesions located at the optic disc overlap with the resulting segmentation mask, and can be removed by this descriptor.

5.2.4 Candidate classification with Random Forest

A Random Forest (RF) is an ensemble classifier that is widely used in the literature due to its capability to perform both classification and feature selection simultaneously [29, 115]. It is also robust against overfitting, which is relevant when having small training sets, and is suitable to deal with noisy, high dimensional imbalanced data. We trained this classifier for the purpose of refining our set of candidates using our hybrid feature vector. In all our experiments, we standardized the features to zero mean an unit variance.

A RF is a combination of T decision trees. These trees are learned from T examples that are randomly sampled with replacement from our training set S. Each node in a tree corresponds to a split made using the best of a randomly selected

Table 5.2: Summary of the hand crafted features used to complement our CNN.

Feature (dimensionality)	**Extracted from**
Average intensity value in the candidate region. (13)	R, G, B, I_W, R_c, G_c, B_c, I_{Wc}, R_{ce}, G_{ce}, B_{ce}, I_{SC},$I_{\text{top-hat}}$
Sum of intensities in the candidate region. (12)	R, G, B, I_W, R_c, G_c, B_c, I_{Wc}, R_{ce}, G_{ce}, B_{ce}, I_{SC}
Standard deviation of intensities in the candidate region. (12)	R, G, B, I_W, R_c, G_c, B_c, I_{Wc}, R_{ce}, G_{ce}, B_{ce}, I_{SC}
Contrast: Difference between mean intensity in the candidate region and mean intensity of the dilated region (12)	R, G, B, I_W, R_c, G_c, B_c, I_{Wc}, R_{ce}, G_{ce}, B_{ce}, I_{SC}
Normalized total intensity: Difference between total and mean intensities of the candidate area in I_{BG}, divided by the candidate's standard deviation in I_{BG}. (3)	G, I_{SC}, I_W
Normalized mean intensity: Difference between mean intensity in I_W and mean intensity of the candidate area in I_{BG}, divided by the standard deviation of the candidate in I_{BG}.(1)	I_W
Minimum intensity in the candidate area. (1)	I_{match}
Area: Number of pixels of the candidate. (1)	BW
Perimeter: Number of pixels on the border of the candidate. (1)	BW
Aspect ratio: Ratio between the major and minor axis lengths. (1)	BW
Circularity = $4\pi\text{Area}/\text{Perimeter}^2$. (1)	BW
Compactness = $\sqrt{(\sum_{j=1}^{n} d_j - \tilde{d})/n}$, where d_j is the distance from the centroid of the object to its jth boundary pixel and $\tilde{d}$ is the mean of all the distances from the centroid to all the edge pixels. n is the number of edge pixels. (1)	BW
Major axis of the ellipse that has the same normalized second central moments as the candidate region. (1)	BW
Minor axis of the ellipse that has the same normalized second central moments as the candidate region. (1)	BW
Eccentricity of the ellipse that has the same second-moments as the candidate region. The eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length. (1)	BW
Ratio of the pixels on the candidate region that are also included in the binary segmentation of the retinal vasculature, obtained as in Section 4.2.1.3. (1)	Vessel segmentation

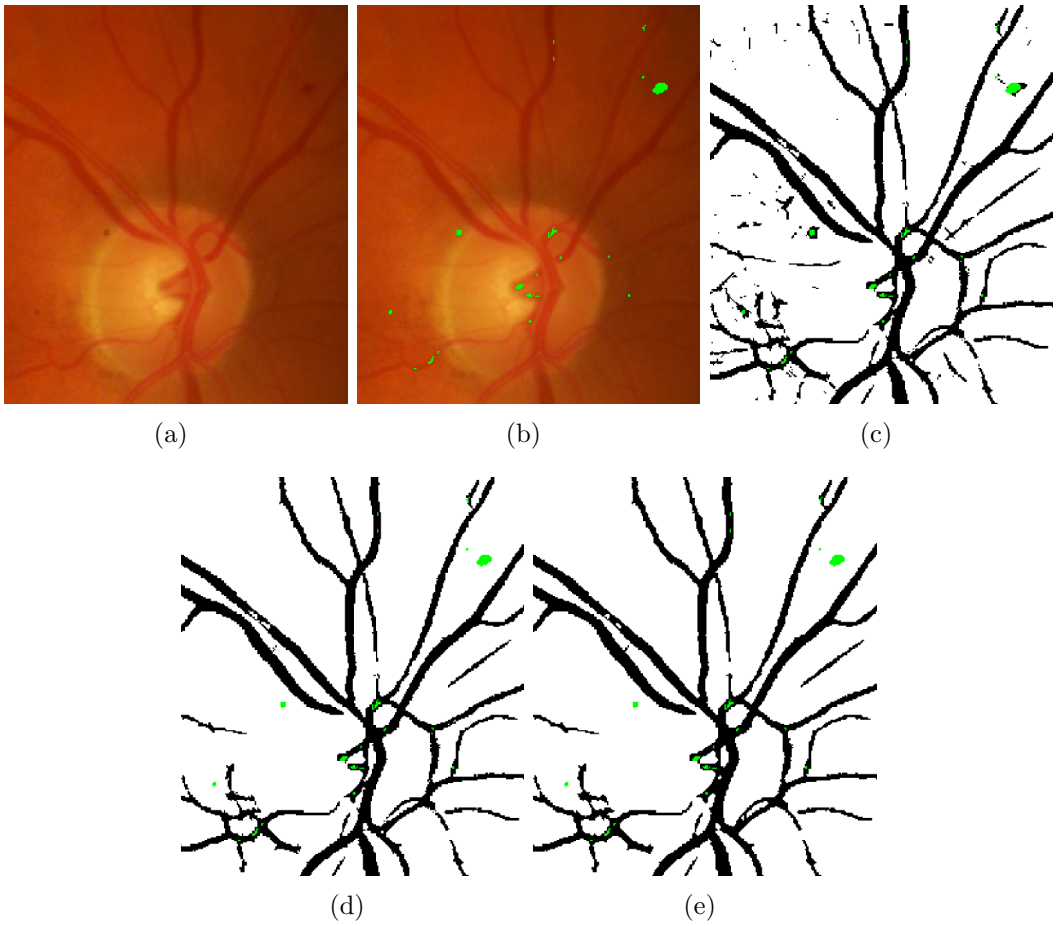


Figure 5.6: Feature based on vessel segmentation. (a) I. (b) I with candidates superimposed. (c) Vessel segmentation. (d) Vessel segmentation after removing spurious elements. (e) Vessel segmentation after morphological closing.

subset of $m = \sqrt{d}$ features, with d being the dimensionality of the feature vector. The quality of the split is given by the decrease in the Gini index that the split produces [29]. Given a feature vector $\mathbf{x}^{(j)}$, the RF evaluates the conditional probability $p_i(c|\mathbf{x}^{(j)})$, where $c \in \{-1, 1\}$ is the class–with -1 corresponding to a non lesion and 1 to a true lesion– and i is the index of the tree in the forest. The final probability is then computed by repeating this process for every tree $0 < i \leq T$, and averaging

Table 5.3: Distribution of DR grades in the MESSIDOR data set, and diagnostic criterion. MA = microaneurysms, HE = hemorrhages and NV= neovascularizations.

Grade	**Criteria**	**Num. images**
R0	$(N_{\mathrm{MA}} = 0)$ AND $(N_{\mathrm{HE}} = 0)$	546
R1	$(0 < N_{\mathrm{MA}} \leq 5)$ AND $(N_{\mathrm{HE}} = 0)$	153
R2	$(5 < N_{\mathrm{MA}} < 15)$ AND $(0 < N_{\mathrm{HE}} < 5)$ AND $(N_{\mathrm{NV}} = 0)$	247
R3	$(N_{\mathrm{MA}} \geq 15)$ OR $(N_{\mathrm{HE}} \geq 5)$ OR $(N_{\mathrm{NV}} > 0)$	254

the responses of each of them:

$$p(c|\mathbf{x}^{(j)}) = \frac{1}{T}\sum_{i}^{T} p_i(c|\mathbf{x}^{(j)}) \tag{5.5}$$

In order to determine the probability P of the image I corresponding to a DR patient or not, we followed the same procedure used by [180]:

$$P(I) = \max_{j} p(c = 1|\mathbf{x}^{(j)}), \tag{5.6}$$

which means that for a given image I with m lesion candidates, the probability of being DR will be associated with the maximum certainty of the classifier of having observed a true positive lesion ($c = 1$).

5.3 Experimental setup

5.3.1 Materials

We conducted experiments using three publicly available data sets: DIARETDB1[1] [97], e-ophtha[2] [48], and MESSIDOR[3] [49]

DIARETDB1 and e-ophtha were used to perform a per-lesion evaluation as they provide lesion level annotations. MESSIDOR provides image level annotations indicating the DR grade, assigned using the criterion detailed in Table 5.3. Thus, this set was used to quantify the performance of our method as a DR screening tool, on

[1] `http://www.it.lut.fi/project/imageret/diaretdb1/`

[2] `http://www.adcis.net/en/Download-Third-Party/E-Ophtha.html`

[3] `http://messidor.crihan.fr`.

a per-image basis. We also used e-ophtha for this purpose, by generating image-level annotations based on the number of red lesions in the ground truth segmentation. Thus, any image with at least one red lesion was labeled as DR. The ROC[4] [142] training set, which comprises 50 fundus photographs taken at different resolutions, was used to augment DIARETDB1 training set for small red lesion detection on e-ophtha. Further details about the experimental setup are provided in Table 5.4.

DIARETDB1 consists of 89 color fundus images taken under varying imaging settings [97]. 84 images contain signs of mild or pre-proliferative DR, and the remaining 5 are considered normal. The entire set is divided into a training set and a test set of 28 and 61 images, respectively. Four different experts have delineated the regions where MA and HE can be found, and a consensus map is provided per each type of lesion. The standard practice is to evaluate MA or HE detection methods at a conservative $\geq 75\%$ agreement [97]. For red lesion detection, however, [180] propose to use as ground truth the union of the consensus maps for both MAs and HEs at a $> 25\%$ level of agreement. We followed this latter approach to evaluate our red lesion detection strategy.

e-ophtha [48] is a database generated from a telemedical network for DR screening, and it includes manual annotations of MAs and small HEs. It comprises 148 images with small red lesions, and 233 with no visible sign of DR. In order to obtain per-image labels indicating the presence or absense of DR, images with any red lesion were labeled as DR.

Finally, MESSIDOR [49] comprises 1200 color fundus images acquired by 3 ophthalmic institutions in France. Images were originally captured at different resolutions, and graded into four different DR stages, being R0 the healthy category and R3 the most severe. Two different classification problems are usually derived from MESSIDOR grades: DR screening, which corresponds to distinguishing R0 from the remaining R1, R2 and R3 grades [16, 180]; and detecting the need for referral, which corresponds to R0 and R1 vs. R2 and R3 grades [162, 176]. We evaluated our method on a per image basis following these two approaches.

Since these data sets do not include FOV masks, which are necessary for processing the images, we automatically generate them by thresholding the luminosity

[4] `http://webeye.ophth.uiowa.edu/ROC/`

Table 5.4: Experimental setup. β is the value for the balanced cross-entropy loss (Equation (5.5)).

Exp. ID	**Detection**	**Training set**	**GT labels**	**True lesions**	**Non lesions**	β	**Per lesion evaluation**	**Per image evaluation**
1	Red lesions with multiple sizes	DIARETDB1 training set (28 images)	MA $>$ 25%$\cup$ HE $>$ 25 %	1059 (27%)	2905 (73%)	$\beta =$ 0.5	DIARETDB1 test set	MESSIDOR
2	Small red lesions	DIARETDB1 & ROC training sets (78 images)	MA $>$ 75% from DIARETDB1 & ROC MA labels	407 (4%)	10282 (96%)	$\beta \sim$ 0.96	e-ophtha	e-ophtha

plane of the CIELab version of the RGB images at 0.15 (for DIARETDB1, e-ophtha and MESSIDOR) and 0.26 (for ROC) [150]. If the resulting binary mask is such that the entire image is estimated as a foreground, an alternative approach is applied where the RGB bands are summed up and the resulting image is thresholded at an empirically tuned value of 150. To smooth borders and reduce noise, all masks are postprocessed with a median filter using square windows of side 5, and only its largest connected component is preserved. The reader can notice that this process is equivalent to the one described in Section 4.3.2. In principle, these masks would be available directly from the fundus camera, and the process of replicating this information directly from the images is a necessary but not central task to the present chapter. The FOV masks for all the data sets used in this chapter are released in the project webpage (Section 5.6).

5.3.2 Model selection

Candidate detection relies on three significant parameters: L, which is the set of scales used to retrieve potential candidates; K, the number of candidates retrieved for a given scale; and px, the minimum area in pixels that a candidate must have. In our experiments, these values were experimentally adjusted using the DIARETDB1 training set, resulting in $L = \{3, 6, 9, \ldots, 60\}$, $K = 120$ and $px = 5$. The maximum scale from L was adapted on the remaining data sets using a scaling factor of $\frac{\mathcal{X}}{1425}$, where 1425 is the average width of the images in DIARETDB1. This allows to recover a set of candidates with a size proportional to the resolution of each image.

The parameters of the CNN (in particular, dropout probability $1-p$ and the size of the fully connected layer N) were designed according to the performance on a held out validation set, randomly sampled from each training set. The parameters that maximized the area under the precision/recall curve ($N = 128$ and $p = 0.99$) were always used for evaluation on the test set. The number of trees $T \in \{100, 110, ..., 200\}$ for the RF was fixed to the value that minimized the out-of-bag error on the training set [29]. The maximum number of possible trees was fixed to a relatively low value (200) to reduce the computational cost during training and prediction. Nevertheless, experiments adding up to 2000 trees to the model did not show any improvements in reducing the out-of-bag error

5.3.3 Evaluation metrics

Free-response ROC (FROC) curves were used to evaluate the performance of our red lesion detection method on a per lesion basis. These plots, which are extensively used in the literature to estimate the overall performance on this task, represent the per lesion sensitivity against the average number of false positive detections per image (FPI) obtained on the data set for different thresholds applied to the candidate probabilities. Thus, FROC curves provide a graphical representation of how the model is able to deal with the detection of true lesions in all the images of the data set. We also computed the Competition Metric (CPM) as proposed in the Retinopathy Online Challenge [142], which is the average per lesion sensitivity at the reference FPI values $\in \{1/8, 1/4, 1/2, 1, 2, 4, 8\}$. The protocol used by [180] was followed when evaluating in DIARETDB1, as indicated in Section 5.3.1.

When evaluating on a per image basis, we used standard ROC curves, where both the sensitivity ($Se = \frac{TP}{FN+TP}$) and $1 -$ specificity ($Sp = \frac{TN}{FP+TN}$) are depicted within the same plot for different DR probability values, obtained as indicated in Equation 5.6. Additionally, we studied the Se at $Sp = 50\%$, which is a standard comparison metric for screening systems [176].

5.4 Results

5.4.1 Per lesion evaluation

Two different experiments were conducted for per lesion evaluation, as detailed in Table 5.4. FROC curves are used for comparison, and Wilcoxon signed rank tests were performed to estimate the statistical significance of the differences in the per lesion sensitivity values. These tests were conducted using 100 sensitivity values retrieved for logarithmically spaced FPI values in the interval $[\frac{1}{8}, ..., 8]$, which corresponds to a more dense version of the reference FPI values used for computing the CPM [142].

Experiment 1 evaluates the model ability to deal with both MAs and HEs simultaneously at multiple scales, following the protocol used in [180] (Figure 5.7). Results of [180] were provided by the authors and obtained using the same training and test configuration, and are included for comparison purposes. Hypothesis tests show a statistically significant improvement in the per lesion sensitivity values when using the combined approach compared to using each representation separately ($p < 2 \times 10^{-18}$ and $p < 4 \times 10^{-17}$ for the CNN probabilities and the hand crafted features, respectively). Moreover, the hybrid method reported better results compared to [180] ($p < 2 \times 10^{-18}$).

As DIARETDB1 includes labels for both MAs and HEs, it is possible to quantitatively assess the accuracy of the method to detect each type of lesion. Figure 5.8 illustrates the FROC curves and the CPM values obtained by the models learned in Experiment 1, when analyzing MAs and HEs separately. For MA detection, the combined approach achieves higher per lesion sensitivity values than using each approach separately ($p < 2 \times 10^{-18}$ and $p < 3 \times 10^{-17}$ for the hand crafted features and the CNN, respectively), with a noticeable improvement at the clinically relevant FPI=1 value (0.2885 versus 0.202 and 0.2 for combined, CNN, and hand crafted, respectively). Moreover, the differences between the manually tuned approach and the CNN probabilities are not statistically significant. When evaluating the ability of the system to detect HEs on the DIARETDB1 test set, it is possible to see that the per lesion sensitivities are higher than those reported for MA detection. Furthermore, the hand crafted features are able to achieve better per lesion sensitivity values than

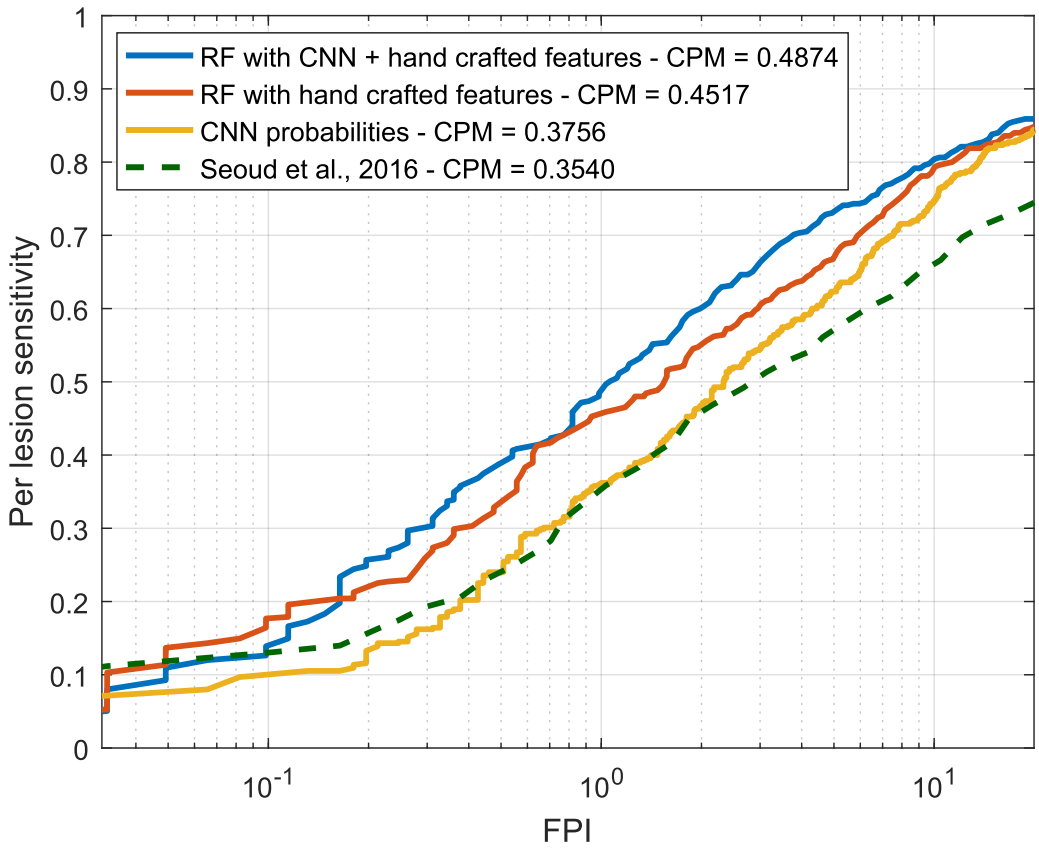
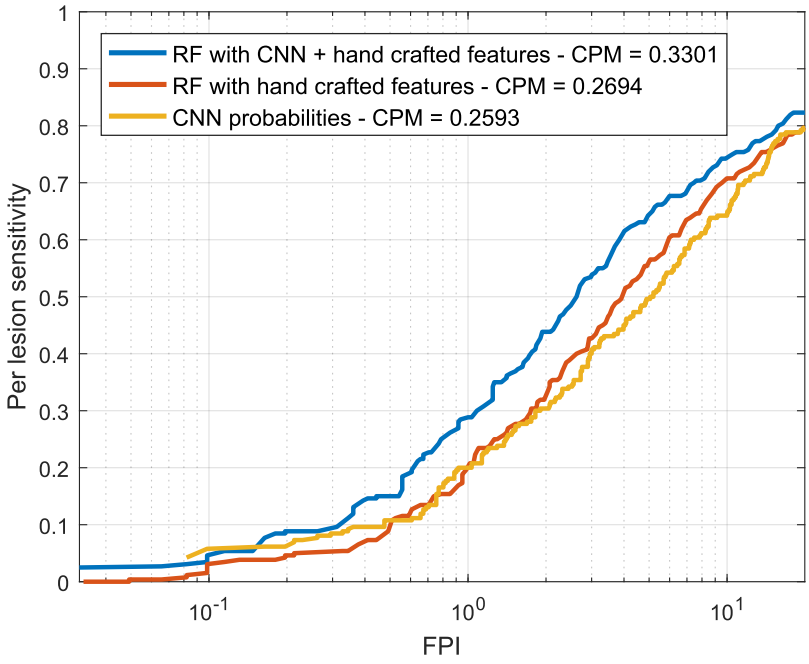


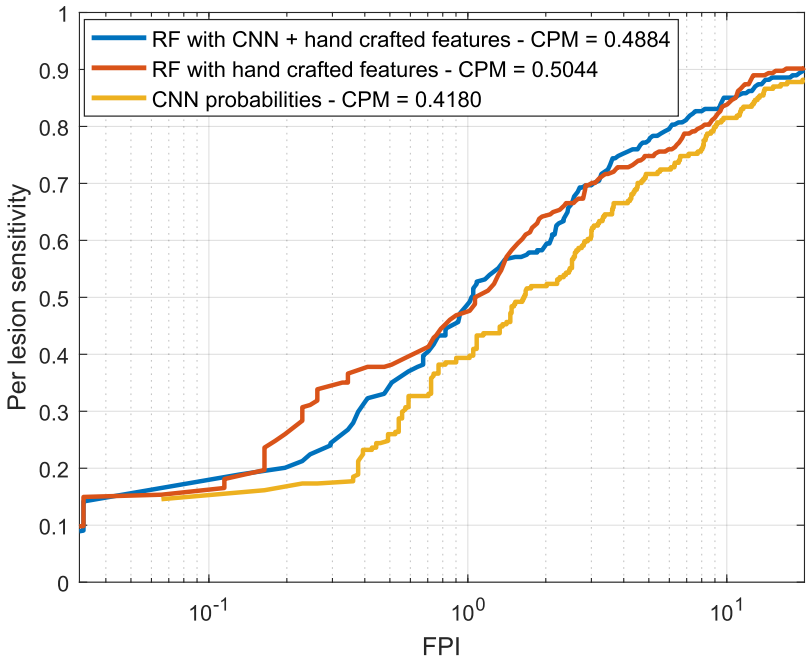
Figure 5.7: Per lesion evaluation in Experiment 1. FROC curve and CPM values obtained on the DIARETDB1 test set.

the combined approach ($p < 5 \times 10^{-5}$) for this specific task. At the clinically relevant FPI value of 1, however, the combined approach reports a slightly higher per lesion sensitivity compared to the manually engineered descriptors (0.4907 versus 0.4724).

Experiment 2 was carried out on e-ophtha to estimate the ability of our method to segment MAs and smaller HEs simultaneously. In this case, a combination of both the DIARETDB1 (MA labels with a level of agreement $\geq$ 75%) and ROC training sets was used for learning, as we observed that few MAs (only 182 for the entire DIARETDB1 set) are retrieved at $\geq$ 75% agreement. To the best of our knowledge, the only method evaluated on e-ophtha is the one presented in [217], although their analysis is performed on a subsample of 74 images with lesions instead of the full data set. By contrast, we used a more challenging evaluation comprising the entire e-ophtha set, including also the 233 images with no visible sign of DR.



(a) Microaneurysms



(b) Hemorrhages

Figure 5.8: Per lesion evaluation for each lesion type in the DIARETDB1 test set.

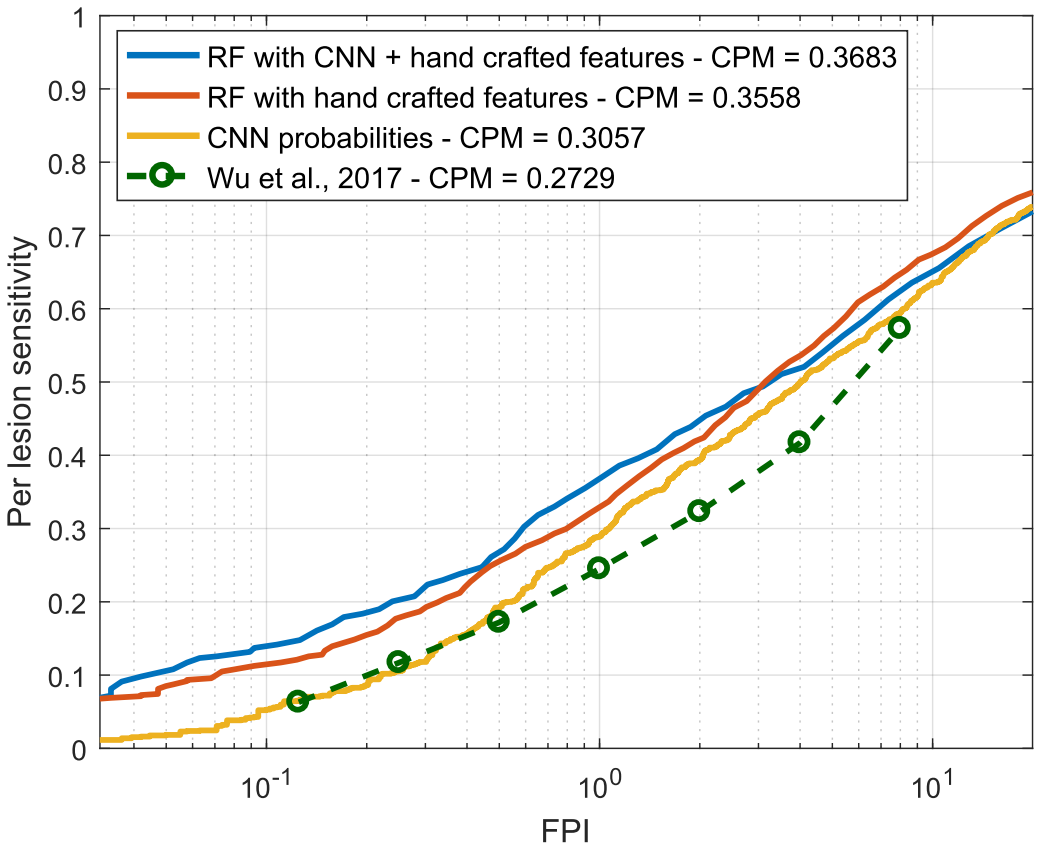


Figure 5.9: Per lesion evaluation in Experiment 2. FROC curve and CPM values obtained on e-ophtha.

Figure 5.9 presents the FROC curves obtained using each approach. As in the previous experiment, the Wilcoxon signed rank tests showed a statistical significant improvement in the per lesion sensitivity values using the hybrid vector of both deep learned features and domain knowledge with respect to the CNN probabilities and the hand crafted features ($p < 2 \times 10^{-18}$ and $p < 2 \times 10^{-9}$, respectively).

Table 5.5 summarizes the CPM values obtained for each experiment and each feature combination, and also using each of the two recently published state-of-the-art methods. Per lesion sensitivities at FPI= 1, which is considered a clinically relevant number of false positives [142] are also provided.

Finally, qualitative results for a randomly selected image in the DIARETDB1 test set are depicted in Figure 5.10. Green circles are detected lesions according to the ground truth labeling provided in the data set, while yellow circles correspond

Table 5.5: CPM values and per lesion sensitivities at FPI= 1 for Experiments 1 and 2.

Method	**Experiment 1**		**Experiment 2**	
	CPM	*Se*	CPM	*Se*
Seoud *et al.*, 2016 [180]	0.3540	0.3462	-	-
Wu *et al.*, 2017 [217]	-	-	0.2729	0.2450
CNN probabilities	0.3756	0.3621	0.3057	0.2894
RF with HCF	0.4517	0.4601	0.3558	0.3291
RF with CNN + HCF	**0.4874**	**0.4883**	**0.3683**	**0.3680**

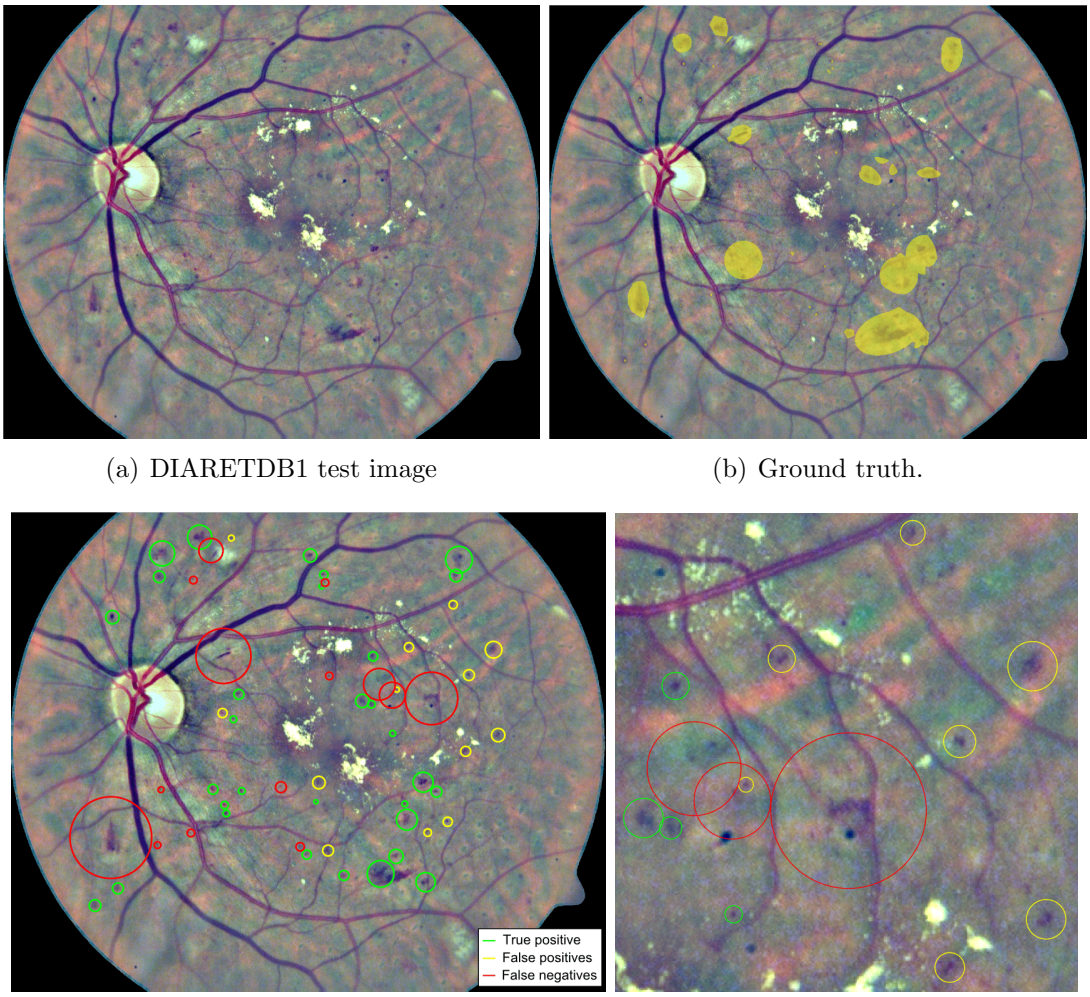
to lesions detected by our method but that are not labeled in the ground truth. Finally, red circles surround the lesions that were manually annotated as true lesions but were ignored by the method. Qualitatively, many of the yellow circles appear to be microaneurysms or hemorrhages that were not detected during manual labeling due to their subtle appearance in the original RGB image.

5.4.2 Per image evaluation

Two different experiments were conducted on MESSIDOR to estimate the performance of our method on a per image basis, one focused on detecting patients with DR, and a second based on detecting those need for immediate referral to a specialist. In both cases, we used the model learned from Experiment 1.

Figure 5.11(a) illustrates the ROC curves for DR screening on MESSIDOR, obtained using our hybrid representation and each of the approaches separately. CNN results were obtained using the network as a classifier. A series of Mann-Whitney U tests ($\alpha = 0.05$) were performed to study the statistical significance of the differences in the AUC values [76]. CNN features (AUC = 0.7912) perform significantly better ($p < 1 \times 10^{-3}$) than hand crafted features (AUC = 0.7325) for this specific task, and the combination of both sources of information results in a substantially higher AUC value of 0.8932 ($p < 1 \times 10^{-6}$). Figure 5.11(b) shows analogous behavior for detecting patients that need referral, with the CNN performing better than the hand crafted features ($p < 2 \times 10^{-3}$), and the combined approach outperforms both individual techniques ($p < 1 \times 10^{-6}$).

Our combined approach shows an analogous behavior when evaluating on e-



(a) DIARETDB1 test image

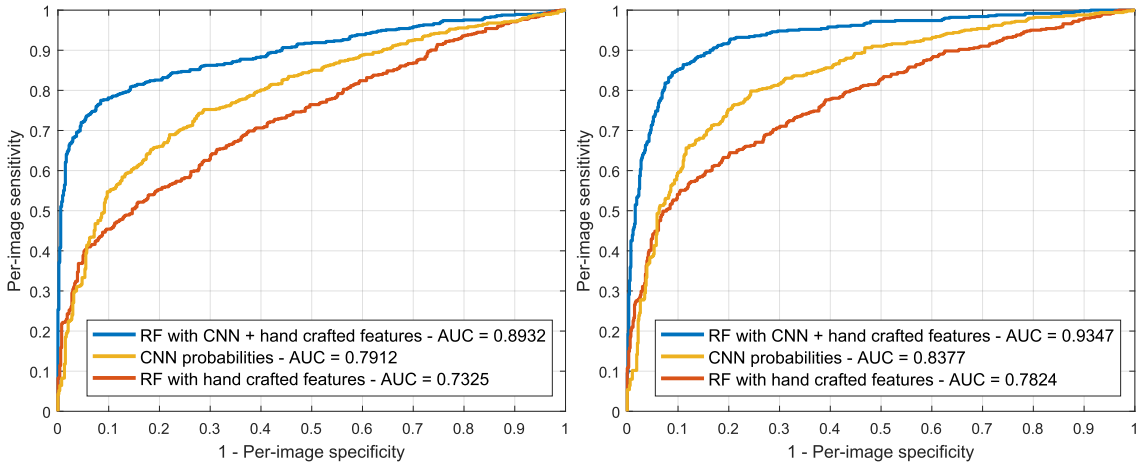
(b) Ground truth.

(c) Red lesion detection

(d) Detail from (c)

Figure 5.10: Qualitative results. (a) image015 from the DIARETDB1 test set. (b) Ground truth labeling at a $> 25\%$ level agreement. (c) Red lesion detections obtained by thresholding the probabilities at 0.644, which corresponds to an average FPI value of 1. (d) Detail from (c) showing lesions unlabeled on the ground truth but identified by our method.

ophtha for DR screening, as illustrated in Figure 5.12. Our combined approach retrieved a significantly higher AUC value (0.9031) than the one reported by the CNN (AUC = 0.8374, $p < 5 \times 10^{-3}$) and the RF classifier trained with hand crafted features (AUC = 0.8812). Hand crafted features perform better than the CNN for screening in this data set, although the difference is not statistically significant



(a) Performance of DR screening

(b) Performance of detecting patients that need referral

Figure 5.11: Per image evaluation. ROC curves for (a) DR screening (R1 vs. R2, R3 and R4) and (b) need for referral (R1 and R2 vs. R3 and R4) on the MESSIDOR data set.

according to the Mann-Whitney U test.

A comparison with respect to other state of the art strategies is presented in Table 5.6. The performance obtained by two human experts, as reported in [176], is also included in the table. The results of the baseline method presented in [180] were obtained using DIARETDB1 as a training set. The other methods included are either based only on red lesion detection or complemented by other features such as image quality assessment or the detection of exudates and neovascularizations.

5.4.3 Feature assessment

In order to assess the visual appearance of the deep learned features, a graphical representation of the 32 filters of size $5 \times 5 \times 3$ learned on the first layer of the CNN is presented in Figure 5.13. From Figure 5.13(a) it is possible to see that filters learned in Experiment 1 are mostly descriptors of the color properties of the lesions. This setting is in line with the fact that the training set used in this case contains not only small MAs but also medium size HEs, which can be more easily described in terms of their internal color homogeneity rather than their edges, which significantly varies from one to another. Other filters are able to capture purple, ellipsoidal structures

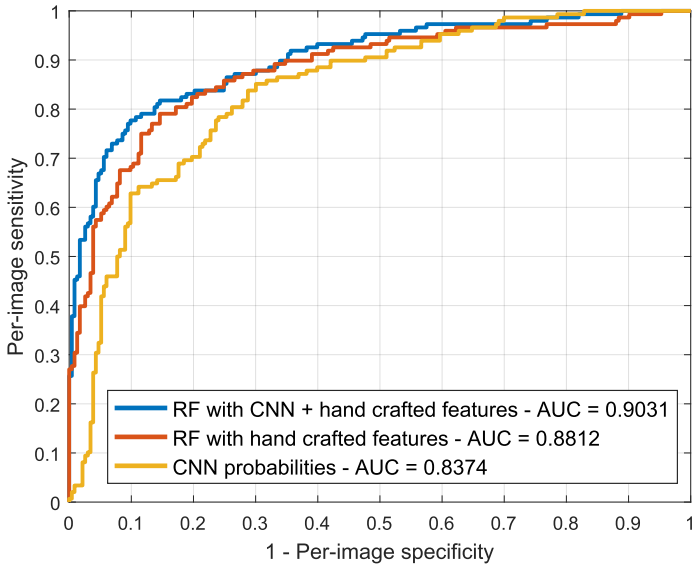


Figure 5.12: Per image evaluation on e-ophtha. ROC curve for DR screening.

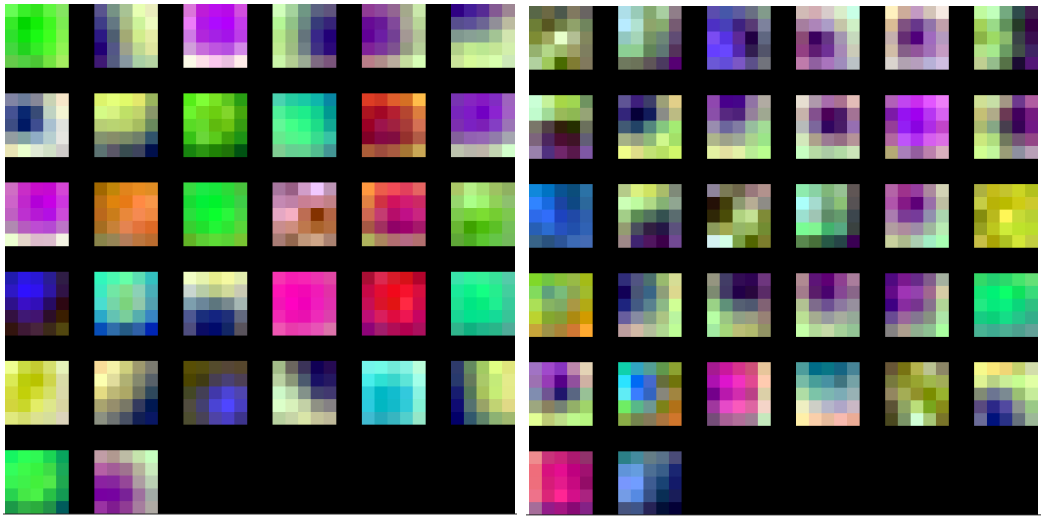
corresponding to true lesions like those illustrated in Figure 5.5(b). This last type of filter is more common in the first layer of the CNN learned in Experiment 2 (Figure 5.13(b)), which might be associated with the smaller true positive structures observed in the training set built with ROC and DIARETDB1 MAs.

The t-distributed stochastic neighbor embedding (t-SNE) method is a recently introduced technique that is widely used to visually assess the ability of multidimensional feature vectors to discriminate different inputs [198]. We followed this approach to study the complementarity of each characterization method, and to qualitatively assess how their integration contribute to improve their original discrimination ability. Figure 5.14 presents the t-SNE mappings of the DIARETDB1 test samples for each characterization approach and for our combined feature vector. The CNN descriptors corresponds to those learned in Experiment 1. The figure also includes a visual representation of the organization of the patches in the embedding space. In general, it is possible to see that the ensemble approach groups the majority of the true positive candidates within a single neighboring area. By contrast, the individual characterization strategies are not able to achieve a single cluster but rather obtain two–in the case of the deep learned features–or more–using the hand crafted features. Detailed regions of the embeddings are depicted in Figure 5.15.

Table 5.6: Comparison of DR screening and need of referral performance on the MESSIDOR data set. Se values correspond to those obtained at a $Sp = 50\%$.

Method	**Screening**		**Need for referral**	
	AUC	**Se**	**AUC**	**Se**
Expert A [176]	0.9220	0.9450	0.9400	0.9820
Expert B [176]	0.8650	0.9120	0.9200	0.9760
Antal and Hajdu, 2012 [16]	0.8750	-	-	-
Costa *et al.*, 2016 [43]	0.8700	-	-	-
Giancardo *et al.*, 2013 [70]	0.8540	-	-	-
Nandy *et al.*, 2016 [134]	-	-	0.9210	-
Pires *et al.*, 2015 [161]	-	-	0.8630	-
Sánchez *et al.*, 2011 [176]	0.8760	**0.9220**	0.9100	0.9440
Seoud *et al.*, 2016 [180] (DIARETDB1)	0.844	-	-	-
Vo and Verma, 2016 [205] (I)	0.8620	-	0.8910	-
Vo and Verma, 2016 [205] (II)	0.8700	-	0.8870	-
HCF	0.7325	0.7645	0.7824	0.8283
CNN	0.7912	0.8471	0.8377	0.9102
HCF + CNN	**0.8932**	0.9109	**0.9347**	**0.9721**

This allows better visualization of particular scenarios such as the patches around the true red lesions, the false positive candidates located in the vascular structures, the artifacts due to speckles of dirt in the lens–which are typical of the images in DIARETDB1–and the false detections within the optic disc. In general, it is possible to observe that CNN features are able to better characterize the orientation and the visual appearance of the true lesion candidates, while the hand crafted features can detect the less obvious lesions under low contrast conditions. The ability of the CNN features to discriminate orientations are more evident when dealing with vascular structures. The hand crafted approach, by contrast, is only able to capture the overall size of the vessels and their intensity properties. When combining both strategies, the main advantages of each of them are maintained. The robustness against artifacts is evident for both the deep learning based and the hand crafted features, as these false positive candidates are grouped together into separate clusters from the true lesions. A similar behavior is observed when dealing with false candidates within the optic disc area.

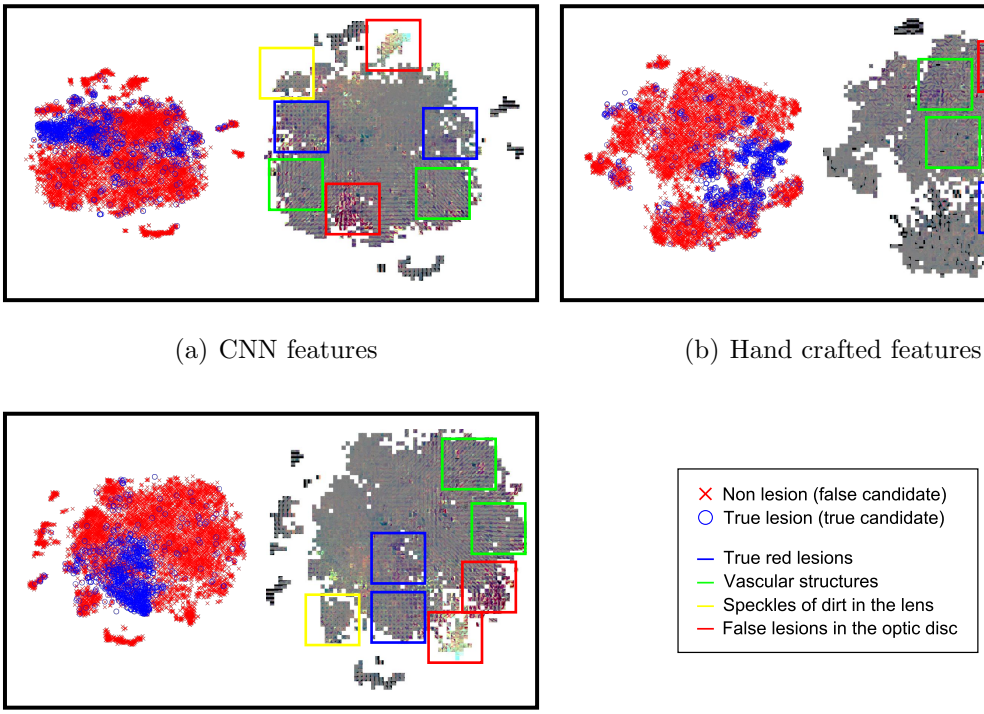


(a) Experiment 1 (b) Experiment 2

Figure 5.13: Learned filters on the first layer of our CNN, as obtained for each experiment in Table 5.4.

5.5 Discussion

In general, the integration of both the deep learned and the hand crafted features significantly improved results compared to using either approach separately. In a per lesion evaluation, the combined approach achieved a consistently higher CPM value both in the e-ophtha and DIARETDB1 test sets, and also a higher per lesion sensitivity for FPI=1, which corresponds to a clinically relevant number of false positives [142]. These values are also higher than those obtained by two recently published baseline methods that were evaluated on the same data set. A similar behavior is observed when evaluating the method on a per image basis. The combined approach improved the performance obtained by each characterization approach separately, meaning that the integration of both sources of information obtains a better characterization of the lesion candidates and, consequently, a more accurate detection of the individual lesions. This is supported by the analysis of the t-SNE mapping obtained for each method (Figures 5.14 and 5.15), in which it is possible to see that the ensemble approach takes advantage of the CNN's ability to characterize fine-grained details such as the orientation of the lesion, while the hand crafted features improve



(a) CNN features

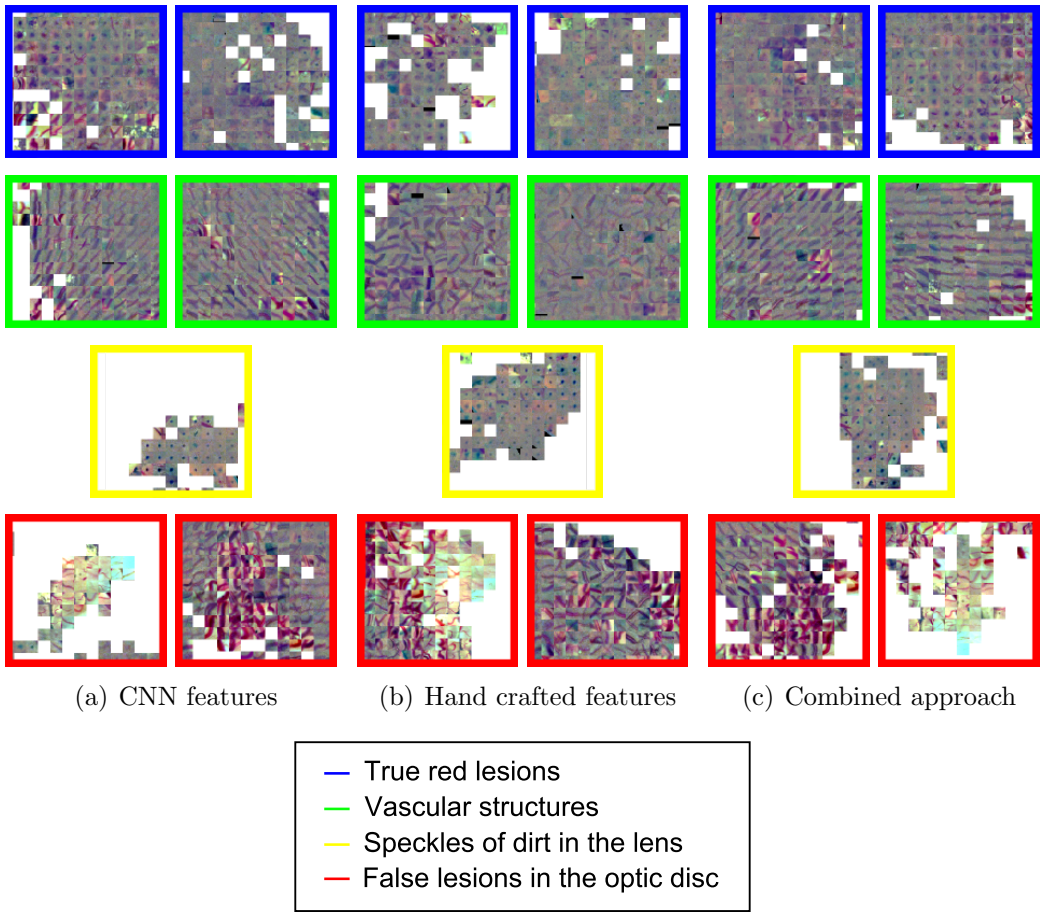
(b) Hand crafted features

(c) Combined approach

Figure 5.14: The t-SNE visualization of the patches from DIARETDB1 test set as mapped using the deep learned features, the hand crafted features and our hybrid feature vector. Left side: color coded labels for each test sample. Right side: patches around the candidates, as visualized using the t-SNE mappings. Details for different types of lesion candidates are shown in Figure 5.15.

the discrimination of low contrast lesions.

Results on the per image evaluation also showed that the proposed strategy is able to achieve higher AUC values than other approaches for DR screening and need-for-referral detection. Moreover, the methods included in Table 5.6 are based not only on red lesion detection [16, 70, 180] but also on additional features such as the assessment of the image quality [176] and/or the presence of other pathological structures such as exudates and neovascularizations [43, 161, 176]. Compared with respect to all these approaches, our method achieved a higher AUC value. Furthermore, it performed better than the DR grading method presented in [205], which uses fine tuned CNNs trained on a data set with 50.000 images with image level annotations. An al-



(a) CNN features (b) Hand crafted features (c) Combined approach

Figure 5.15: Details from the *t*-SNE visualization in Figure 5.14 for different types of red lesion candidates (true lesions, vascular structures, speckles of dirt in the lens and false detections in vessel curves in the optic disc).

most equal performance was obtained for DR screening compared with the recently published method by Quellec *et al.* [166], which reported an AUC= 0.893 in the MESSIDOR data set. However, such an approach uses multiple images per patient, contextual information and clinical records to learn diagnostic rules from a data set with 12.000 examinations. Our method is able to achieve a slightly higher AUC value without including any additional clinical information. Furthermore, a competitive Se value was obtained in comparison with Expert B [176], indicating that this approach can match the ability of a human observer for DR screening and detecting patients that need referral. Thus, our automated red lesion detection system could

be integrated in a more general DR screening platform to improve the ability to detect DR patients. Furthermore, incorporating other modules for detecting other pathological structures can eventually improve the reported performance.

It is also important to underline that all the stages in the proposed method have parameters that are automatically adjusted to each image resolution. Their values, which are reported in Section 5.3.2, were empirically selected using different data than that used for evaluation, and were proportionally scaled in the subsequent experiments to compensate resolution changes. This simple approach provides an approximate scale invariance that is valuable to facilitate the adaptability of the method to be applied on images obtained using different fundus cameras.

When analyzing each individual characterization approach, it is possible to see in Experiment 1 that both the RF trained with hand crafted features and the CNN achieved higher per lesion sensitivities than the method described in [180] ($p < 2\times10^{-18}$ and $p < 2\times10^{-4}$, respectively). This is likely due to the fact that our method for extracting candidates differs from the one used by the alternative approach. Moreover, [180] eliminate the lesion candidates occurring within an estimated area around the optic disc center, which is determined using an automated approach. As a consequence, if the diameter of the optic disc is accidentally overestimated by such a method, candidates within valid regions will be suppressed and it will not be possible to recover them afterwards during the classification stage. As seen in Figures 5.14 and 5.15, our combined approach is able to discriminate the candidates within the optic disc area and the vascular structures. Hence, instead of using a rigid elimination step based on optic disc segmentation, we let the classifier to decide whether a candidate is actually a true positive or a false positive occurring on an anatomical region. This approach increases the maximum achievable per lesion sensitivity on each image, allowing to train our classifier with a larger amount of false positive lesions and to get a higher sensitivity in test time. A similar observation can be made from the results of Experiment 2, in which the hand crafted features and the deep learning based approach reported higher per lesion sensitivities than those reported by [217]. It must be underlined, also, that the method in [217] was trained on the first half of the images with pathologies on e-ophtha and evaluated on the second half, rather than trained on a separate data set and evaluated on

the complete set, as in our case. Moreover, it is worth noting that the images of the healthy patients were also included during evaluation to get a more accurate estimation of its actual performance on a real, clinical scenario.

On a per image basis, it is possible to see that the individual approaches trained in Experiment 1 are not able to achieve AUC values higher than those reported in [180] (Table 5.6). This is likely due to the fact that, as indicated by the authors, their method is more accurate for detecting blot HEs and MAs than HEs with other shapes. The images in MESSIDOR were originally graded as R0 and R1 taking into account the number of MAs (Table 5.3) [49]. Hence, being more accurate in the detection of MAs will result in a better ability to distinguish much earlier stages. When individually using the hand crafted features or the CNN, both methods are less precise for detecting MAs but better for discriminating other HEs. This argument is supported by results presented in Figure 5.8, in which it is possible to see that the per lesion sensitivity values obtained for MA detection are lower than those reported for HEs. Moreover, it was observed that the CNN performed equally or better than the RF trained with manually engineered features on the low FPI regime for MA detection. This explains the behavior observed in Figure 5.11, where the CNN probabilities achieved a higher AUC value for DR screening and need for referral detection. Nevertheless, the combination of both approaches with the RF classifier consistently improved their individual performance, achieving a much better characterization of the MAs (as observed in the improvements reported in Figure 5.8(a)) and, consequently, a better discrimination of the DR patients.

5.6 Conclusions

We have proposed a novel method for red lesion detection in fundus images based on a hybrid vector of both CNN-based and hand crafted features. A CNN is trained using patches around lesion candidates to learn features automatically, and those descriptors are complemented using domain knowledge to improve their discrimination ability. Results on benchmark data sets empirically demonstrated that the resulting system achieves a new state-of-the-art in this domain, and that combining both sources of information provides statistically significant improvements compared

to using each of them separately. A similar behavior is observed when evaluating our screening system both for DR and need-for-referral detection, reporting higher AUC values than those obtained by other existing approaches based not only on red lesion detection but also on analyzing other pathologies such as bright lesions or neovascularizations, or even learning classifiers using additional clinical information. Considering the high cost of manually labeling fundus photographs at a lesion level, our method represents a robust alternative to improve performance of other deep learning based approaches.

Chapter 6

Conclusions and future lines of research

6.1 Contributions

In this thesis we have presented several contributions for automated fundus image analysis, with applications on the computer-aided diagnosis of DR and glaucoma. These two diseases are the main causes of preventable blindness in the world, and their detection using fundus images is valuable due to the low cost and the non invasive nature of this imaging modality.

In Chapter 3 we have presented a novel retinal vessel segmentation method based on a fully connected conditional random field model (FC-CRF), discriminatively trained using a structured output support vector machine (SOSVM). In general, shape priors are difficult to be introduced in supervised classifiers. Moreover, standard priors such as a Potts model or total variation usually fail to deal with elongated structures. We empirically demonstrated that fully connected pairwise potentials are more expressive, allowing to recover more accurate representations of the retinal vasculature that includes thin vessels that are otherwise ignored by conventional local neighborhood based approaches. The fully connected model was learned using a SOSVM, a task that used to be unfeasible before due to the high complexity of the inference in fully connected graphs. By using the energy model and the inference approach proposed in [102] we were able to train the FC-CRF so that the

weights for the unary and the pairwise potentials are learned simultaneously. An extensive evaluation and comparison using benchmark data sets such as DRIVE, STARE, CHASEDB1 and HRF showed that our method outperforms several existing approaches, achieving state of the art performance as measured in terms of global accuracy measures such as the Matthews Correlation Coefficient, the F1-score and the G-mean. Moreover, this method was also used by the other methods proposed in this thesis, showing a significant ability to deal not only with images of healthy patients but also for images of patients affected by DR and glaucoma, requiring a minimum of postprocessing to remove artifacts on the lesions.

A novel strategy for glaucoma screening based on transferring convolutional neural networks (CNNs) was introduced in Chapter 4. As only few, small data sets are publicly available for designing automated methods for glaucoma detection, training CNNs from scratch is not feasible yet for this task. To mitigate this issue, we proposed to characterize patients with and without the disease using features retrieved by a pre-trained from non-medical data CNN. This network was fed with images adjusted using state of the art preprocessing methods. Our experiments showed a competitive performance compared to other strategies that were evaluated on the same benchmark data set, DRISHTI-GS1. Moreover, we empirically observed that removing the retinal vasculature using image inpainting significantly reduced the performance of the CNN features, which might indicate that the vessels provide valuable information to perform the diagnostics.

Finally, in Chapter 5 we have presented a hybrid approach for DR screening based on red lesion detection. Current methods for computer-aided diagnosis of DR are aided by the detection of microaneurysms (MAs) and small hemorrhages (HEs), which, despite being the earliest signs of the disease, require an intensive effort to be found manually. We observed that current approaches for detecting these lesions are based on a two stage approach consisting first on detecting a set of lesion candidates and then removing the false positives by means of a supervised classifier trained with hand crafted features. We have proposed to improve this existing pipeline by adding deep learned features from a CNN trained from scratch on candidates patches. As the size of the available data sets is limited, we used a relatively shallow architecture to avoid overfitting. The ensemble feature vector of both hand crafted and deep learned

features was used to train a Random Forest (RF) classifier, which is known to perform well under class inbalance and high dimensional data. Our experimental validation on publicly available data sets of fundus images such as DIARETDB1, e-ophtha and MESSIDOR showed that this approach is able to achieve a consistently high performance when evaluated both on a per-lesion and a per-image basis. Moreover, we achieved higher performance for DR screening compared with the opinion of a second human observer.

6.2 Future lines of research

Several lines of research can be derived from the contributions made in this thesis. We will discuss some of them in the following sections.

6.2.1 Vessel segmentation and characterization

Our vessel segmentation method is able to achieve accurate representations of the retinal vasculature on images of healthy subjects, glaucomatous patients or early DR cases. When dealing with advanced DR subjects, however, false positives might occur in the presence of large hemorrhages or exudates. We partially solved this issue by applying postprocessing techniques as those described in Chapter 5, although this is done at the cost of losing some of the thinner vessels. This artifacts are due to the nature of the hand crafted features that we used, which are known to be affected by the darkest lesions or the dark streak patterns that exist in between bright lesions, for example. A valuable option to mitigate this issue would be to first segment out these structures, and then use image inpainting to remove them from the original images, before computing the vessel segmentation. This approach was followed in [14], where exudates are inpainted to improve the discrimination ability of Hessian based features. This approach might be revisited in the context of our segmentation approach to evaluate if it can potentially contribute to remove these false detections. Another alternative is to train deep neural networks and use the learned filters as an input for a FC-CRF. This idea was recently taken by other authors, motivated by the results we presented here. In [66, 67], Fu and colleagues proposed to simultaneously train a FC-CRF and a CNN, achieving better

performance on DRIVE and STARE, under the existence of pathological structures.

Another issue related with the hand crafted features is that they require an intensive effort to be adjusted to other image resolutions. This difficulty might be overcome by using a different approach for scaling their parameters, such as the one recently proposed in [206]. By means of a simple learning approach such as linear regression, different curves can be designed for parameter scaling. Deep learning based approaches also suffer from this issue, which explains why these techniques are usually evaluated only on low resolution data sets such as DRIVE and STARE. Recalling that the major difficulties on vessel segmentation is the detection of the narrower vessels, which are more evident at higher resolutions, it is necessary to start to focus on data sets such as HRF or the recently introduced DRHAGIS [82].

A still open problem on retinal image analysis is the automated classification of arteries and veins (A/V classification), which is valuable for the characterization of vascular pathological processes associated to different diseases [2]. This task is challenging as the visual differences between the two classes are not evident. Several methods have been proposed to characterize such differences by means of hand crafted features. One potential alternative would be to combine these features and filters for vessel detection to perform both segmentation and classification of vessels into arteries and veins simultaneously. As SOSVM and FC-CRFs are sufficiently general to be used both for binary and multiclass classification, our method for blood vessel segmentation can be easily adapted for this new task. One potential drawback of this alternative relies on the hand crafted features for A/V classification, which might be not sufficient to discriminate these two classes. This issue might be partially mitigated using also deep learned features.

6.2.2 Automated glaucoma screening

Glaucoma detection is still challenging and extremely required due to the large number of undiagnosed patients [164]. Currently there are different initiatives pushing to apply deep learning techniques for detecting this disease from optical coherence tomography (OCT) scans, such as the partnership between Google DeepMind and

Moorfields Eye Hospital.[1] However, less interest has been observed for applying this family of methods on fundus images for glaucoma detection. We still believe that this approach has to be exploited, specially taking into account the low cost of the acquisition device and its potential to be applied on screening campaigns. Its main complication is related with the lack of large scale data sets to train deep learning models, however. Current databases are mostly designed for ONH segmentation, for instance, or have been labelled based only on the assessment of the optic disc. As deep learning techniques are able to learn the most discriminative image properties by themselves, it is extremely necessary to use as much information as possible at the time of assigning the ground truth labels for the training sets. If not, then the CNNs are likely to only learn how to measure the cup-to-disc ratio, for example, which is valuable but still not enough for an accurate glaucoma diagnosis [75].

Our method for glaucoma detection has been demonstrated to be as accurate as other hand tuned approaches by using features obtained using an off-the-shelf CNN. If large scale data sets were available, then an immediate continuation of this work would be to fine tune the network weights by initializing the CNN parameters with the pre-trained network, and resuming the optimization process. Another potential strategy would be to train a CNN from scratch using other data sets of fundus images that were designed for different diseases, such as the one from the Kaggle Challenge on DR detection[2], and then fine tune those weights for glaucoma detection. This configuration might accelerate the convergence of the CNN, as the already learned filters might be more associated with the properties of the fundus images than those learned from natural images.

Finally, an interesting work that might also be derived from our contribution is the development of an ensemble method for glaucoma detection based on multi-source information. The probability given by our logistic regression classifier can be used as part of a feature vector composed of other interesting parameters such as the intraocular pressure (IOP) or the patient clinical record, for example. This approach might be able to capture other representative patterns that can be useful to achieve more accurate diagnostics.

[1] `https://www.theguardian.com/technology/2016/jul/05/google-deepmind-nhs-machine-learning-blindness`

[2] `https://kaggle.com/c/diabetic-retinopathy-detection`

6.2.3 Automated DR screening

Our red lesion detection method has a direct impact in the development of computer-assisted diagnosis tools for DR screening, as it has demonstrated its ability to recognize the existence of such lesions with a significant accuracy. However, there is still room for further improvements. The integration with similar techniques but for detecting bright lesions such as exudates or cotton-wool spots is extremely necessary, as these pathological structures are also relevant on a clinical setting.

For automated DR screening, we observed that the red lesion detection approach that we presented in this thesis is able to achieve high performance too. DR grading, however, is a more challenging task, as it is necessary not only to detect MAs or HEs but also to determine if other lesions such as exudates or neovascularizations occurs. One potential improvement to our system would be to train a CNN using the Kaggle challenge data set, which is composed of more than 90.000 fundus photographs, and integrates its output with the output of our system. Recently, different authors have presented promising results on this data set [73]. However, as CNNs usually work on downsized versions of the original images, they are usually able to identify large lesions but no MAs or small HEs, which, as we previously saw in Chapter 5, are subtle and small. This setting makes important to combine the red lesion detection with the CNN output.

Alternatively, features based on the probability distribution of the red lesion candidates can also be used for DR grading. In the MESSIDOR data set, for instance, the DR grades were assigned by counting the number of MAs and HEs, and also assessing the presence of neovascularizations. Using other features such as histograms representing the distribution of the outputs of our Random Forest classifier, for example, might reduce the effect that the outliers in the red lesion detection step introduces in the automated diagnosis. Furthermore, an approach similar to the one we followed for glaucoma detection can be taken into account for characterizing mild or pre-proliferative cases, in which large HEs affect the overall distribution on the image intensities and, as a consequence, the outputs of the pre-trained CNN. Finally, vessel characterization techniques such as the analysis of the fractal dimension of the vasculature could also be integrated for detecting the proliferative cases.

Bibliography

[1] A*. SIVA brochure. https://www.etpl.sg/qql/slot/u94/Software%20to%20License/SIVA/SIVA%20ebrochure.pdf. Accessed: 2017-03-23.

[2] M. D. Abràmoff, M. K. Garvin, and M. Sonka. Retinal imaging and image analysis. *IEEE Reviews in Biomedical Engineering*, 3:169–208, 2010.

[3] M. D. Abràmoff and M. Niemeijer. Mass screening of diabetic retinopathy using automated methods. In *Teleophthalmology in Preventive Medicine*, pp. 41–50. Springer, 2015.

[4] R. Acharya, C. K. Chua, E. Ng, W. Yu, and C. Chee. Application of higher order spectra for the identification of diabetes retinopathy stages. *Journal of Medical Systems*, 32(6):481–488, 2008.

[5] U. R. Acharya, E. Y.-K. Ng, J.-H. Tan, S. V. Sree, and K.-H. Ng. An integrated index for the identification of diabetic retinopathy stages using texture parameters. *Journal of Medical Systems*, 36(3):2011–2020, 2012.

[6] M. Agudelo-Botero and C. A. Dávila-Cervantes. Carga de la mortalidad por diabetes mellitus en América latina 2000-2011: los casos de Argentina, Chile, Colombia y México. *Gaceta Sanitaria*, 29(3):172–177, 2015.

[7] C. Agurto, S. Murillo, V. Murray, M. Pattichis, S. Russell, M. Abramoff, and P. Soliz. Detection and phenotyping of retinal disease using AM-FM processing for feature extraction. In *42nd Asilomar Conference on Signals, Systems and Computers*, pp. 659–663. IEEE, 2008.

[8] B. Al-Diri, A. Hunter, and D. Steel. An active contour model for segmenting and measuring retinal vessels. *IEEE Transactions on Medical Imaging*, 28(9):1488–1497, 2009.

[9] B. Al-Diri, A. Hunter, D. Steel, and M. Habib. Automated analysis of retinal vascular network connectivity. *Computerized Medical Imaging and Graphics*, 34(6):462–470, 2010.

[10] M. Al-Rawi, M. Qutaishat, and M. Arrar. An improved matched filter for blood vessel detection of digital retinal images. *Computers in Biology and Medicine*, 37(2):262–267, 2007.

[11] N. Amerasinghe, T. Aung, N. Cheung, C. W. Fong, J. J. Wang, P. Mitchell, S.-M. Saw, and T. Y. Wong. Evidence of retinal vascular narrowing in glaucomatous eyes in an Asian population. *Investigative Ophthalmology & Visual Science*, 49(12):5397–5402, 2008.

[12] M. A. Amin and H. Yan. High speed detection of retinal blood vessels in fundus image using phase congruency. *Soft Computing*, 15(6):1217–1230, 2011.

[13] M. M. Angelica, A. Sanseau, and C. Argento. Arterial narrowing as a predictive factor in glaucoma. *International Ophthalmology*, 23(4):271–274, 2001.

[14] R. Annunziata, A. Garzelli, L. Ballerini, A. Mecocci, and E. Trucco. Leveraging multiscale hessian-based enhancement with a novel exudate inpainting technique for retinal vessel segmentation. *IEEE Journal of Biomedical and Health Informatics*, 20(4):1129–1138, 2016.

[15] R. Annunziata and E. Trucco. Accelerating convolutional sparse coding for curvilinear structures segmentation by refining SCIRD-TS filter banks. *IEEE Transactions on Medical Imaging*, 35(11):2381–2392, 2016.

[16] B. Antal and A. Hajdu. An ensemble-based system for microaneurysm detection and diabetic retinopathy grading. *IEEE Transactions on Biomedical Engineering*, 59(6):1720–1726, 2012.

[17] O. Arend, A. Remky, N. Plange, B. Martin, and A. Harris. Capillary density and retinal diameter measurements and their impact on altered retinal circulation in glaucoma: a digital fluorescein angiographic study. *British Journal of Ophthalmology*, 86(4):429–433, 2002.

[18] A. Argyriou, R. Foygel, and N. Srebro. Sparse prediction with the k-support norm. In *Advances in Neural Information Processing Systems*, pp. 1457–1465. 2012.

[19] E. Arlington. http://www.eyeglasses-arlington.com/diabetic-retinopathy.html. `http://www.eyeglasses-arlington.com/diabetic-retinopathy.html`. Accessed: 2017-03-23.

[20] G. Azzopardi, N. Strisciuglio, M. Vento, and N. Petkov. Trainable COSFIRE filters for vessel delineation with application to retinal images. *Medical Image Analysis*, 19(1):46–57, 2015.

[21] P. Bankhead, C. N. Scholfield, J. G. McGeown, and T. M. Curtis. Fast retinal vessel detection and measurement using wavelets and edge location refinement. *PLOS ONE*, 7(3):e32435, 2012.

[22] Y. Bar, I. Diamant, L. Wolf, and H. Greenspan. Deep learning with non-medical training used for chest pathology identification. In *SPIE Medical Imaging*, pp. 94140V–94140V. International Society for Optics and Photonics, 2015.

[23] R. Barrenechea, I. d. la Fuente, R. G. Plaza, N. Flores, L. Segovia, Z. Villagómez, E. E. Camarero, L. C. Zepeda-Romero, V. C. Lansingh, H. Limburg, et al. Encuesta nacional de ceguera y deficiencia visual evitable en Argentina, 2013. *Pan American Journal of Public Health*, 37(1), 2015.

[24] E. J. Bekkers, J. Zhang, R. Duits, and B. M. ter Haar Romeny. Curvature based biomarkers for diabetic retinopathy via exponential curve fits in SE(2). In E. Trucco, X. Chen, Garvin, M. K., J. J. Liu, and X. Y. Frank, eds., *Proceedings of the Ophthalmic Medical Image Analysis Second International Workshop, OMIA*. 2015.

[25] Y. Bengio, I. J. Goodfellow, and A. Courville. Deep learning. *Nature*, 521:436–444, 2015.

[26] R. Bock, J. Meier, G. Michelson, L. Nyúl, and J. Hornegger. Classifying glaucoma with image-based features from fundus photographs. *Pattern Recognition*, pp. 355–364, 2007.

[27] R. Bock, J. Meier, L. G. Nyúl, J. Hornegger, and G. Michelson. Glaucoma risk index: automated glaucoma detection from color fundus images. *Medical Image Analysis*, 14(3):471–481, 2010.

[28] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004.

[29] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[30] A. Budai et al. Robust vessel segmentation in fundus images. *International Journal of Biomedical Imaging*, 2013, 2013.

[31] T. Chakraborti, D. K. Jha, A. S. Chowdhury, and X. Jiang. A self-adaptive matched filter for retinal blood vessel detection. *Machine Vision and Applications*, 26(1):55–68, 2014.

[32] A. Chakravarty and J. Sivaswamy. Glaucoma classification with a fusion of segmentation and image-based features. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, volume 1, pp. 689–692. IEEE, 2016.

[33] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014. doi: http://dx.doi.org/10.5244/C.28.6.

[34] X. Chen, Y. Xu, D. W. K. Wong, T. Y. Wong, and J. Liu. Glaucoma detection based on deep convolutional neural network. In *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 715–718. IEEE, 2015.

[35] X. Chen, Y. Xu, S. Yan, D. W. K. Wong, T. Y. Wong, and J. Liu. Automatic feature learning for glaucoma detection based on deep learning. In *Medical Image Computing and Computer-Assisted Intervention-MICCAI*, pp. 669–677. Springer, 2015.

[36] E. Cheng, L. Du, Y. Wu, Y. J. Zhu, V. Megalooikonomou, and H. Ling. Discriminative vessel segmentation in retinal images by fusing context-aware hybrid features. *Machine Vision and Applications*, 25(7):1779–1792, 2014.

[37] C. Y.-l. Cheung, E. Lamoureux, M. K. Ikram, M. B. Sasongko, J. Ding, Y. Zheng, P. Mitchell, J. J. Wang, and T. Y. Wong. Retinal vascular geometry in asian persons with diabetes and retinopathy. *Journal of Diabetes Science and Technology*, 6(3):595–605, 2012.

[38] N. Cheung, K. C. Donaghue, G. Liew, S. L. Rogers, J. J. Wang, S.-W. Lim, A. J. Jenkins, W. Hsu, M. L. Lee, and T. Y. Wong. Quantitative assessment of early diabetic retinopathy using fractal analysis. *Diabetes Care*, 32(1):106–110, 2009.

[39] F. Ciompi, B. de Hoop, S. J. van Riel, K. Chung, E. T. Scholten, M. Oudkerk, P. A. de Jong, M. Prokop, and B. van Ginneken. Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. *Medical Image Analysis*, 26(1):195–202, 2015.

[40] T. A. Ciulla, A. G. Amador, and B. Zinman. Diabetic retinopathy and diabetic macular edema. *Diabetes Care*, 26(9):2653–2664, 2003.

[41] V. L. Clark and J. A. Kruse. Clinical methods: the history, physical, and laboratory examinations. *JAMA*, 264(21):2808–2809, 1990.

[42] W. Commons. Schematic diagram of the human eye: Wikimedia commons. `https://commons.wikimedia.org/wiki/File:Schematic_diagram_of_the_human_eye_en.svg`. Accessed: 2017-03-23.

[43] J. Costa, I. Sousa, and F. Soares. Smartphone-based decision support system for elimination of pathology-free images in diabetic retinopathy screening.

In *International Conference on IoT Technologies for HealthCare*, pp. 83–88. Springer, 2016.

[44] R. Crosby-Nwaobi, L. Z. Heng, and S. Sivaprasad. Retinal vascular calibre, geometry and progression of diabetic retinopathy in type 2 diabetes mellitus. *International Journal of Ophthalmology*, 228(2):84–92, 2011.

[45] J. Cuadros and G. Bresnick. EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. *Journal of Diabetes Science and Technology*, 3(3):509–516, 2009.

[46] P. Dai, H. Luo, H. Sheng, Y. Zhao, L. Li, J. Wu, Y. Zhao, and K. Suzuki. A new approach to segment both main and peripheral retinal vessels based on gray-voting and gaussian mixture model. *PLOS ONE*, 10(6):e0127748, 2015.

[47] D. C. DeBuc. The role of retinal imaging and portable screening devices in tele-ophthalmology applications for diabetic retinopathy management. *Current Diabetes Reports*, 16(12):132, 2016.

[48] E. Decencière, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno, et al. Teleophta: Machine learning and image processing methods for teleophthalmology. *IRBM*, 34(2):196–203, 2013.

[49] E. Decencière, X. Zhang, G. Cazuguel, B. Laÿ, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, et al. Feedback on a publicly distributed image database: the Messidor database. *Image Analysis & Stereology*, 33(3):231–234, 2014.

[50] S. Dua, U. R. Acharya, P. Chowriappa, and S. V. Sree. Wavelet-based energy features for glaucomatous image classification. *IEEE Transactions on Information Technology in Biomedicine*, 16(1):80–87, 2012.

[51] B. Dupas, T. Walter, A. Erginay, R. Ordonez, N. Deb-Joardar, P. Gain, J.-C. Klein, and P. Massin. Evaluation of automated fundus photograph analysis algorithms for detecting microaneurysms, haemorrhages and exudates, and of a

computer-assisted diagnostic system for grading diabetic retinopathy. *Diabetes & Metabolism*, 36(3):213–220, 2010.

[52] M. Esmaeili et al. A new curvelet transform based method for extraction of red lesions in digital color retinal images. In *17th IEEE International Conference on Image Processing (ICIP)*, pp. 4093–4096. IEEE, 2010.

[53] L. Espona, M. J. Carreira, M. Penedo, and M. Ortega. Retinal vessel tree segmentation using a deformable contour model. In *19th International Conference on Pattern Recognition*, pp. 1–4. IEEE, 2008.

[54] C. F. Etienne. Reducing avoidable blindness and visual impairment in the region of the Americas. *Revista Panamericana de Salud Pública*, 37(1):1–3, 2015.

[55] K. Falconer. *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2004.

[56] A. Fathi and A. R. Naghsh-Nilchi. Automatic wavelet-based retinal blood vessels segmentation and vessel diameter estimation. *Biomedical Signal Processing and Control*, 8(1):71–80, 2013.

[57] O. Faust, R. Acharya, E. Y.-K. Ng, K.-H. Ng, and J. S. Suri. Algorithms for the automated detection of diabetic retinopathy using digital fundus images: a review. *Journal of Medical Systems*, 36(1):145–157, 2012.

[58] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[59] A. Fercher and E. Roth. Ophthalmic laser interferometry. *SPIE Milestone Series*, 165:242–245, 2001.

[60] F. Fiorani and U. Schurr. Future scenarios for plant phenotyping. *Annual Review of Plant Biology*, 64:267–291, 2013.

[61] M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, C. Owen, A. Rudnicka, and S. Barman. Retinal vessel extraction using first-order derivative of gaussian

and morphological processing. In *Advances in Visual Computing*, pp. 410–420. Springer, 2011.

[62] M. M. Fraz, A. Basit, and S. Barman. Application of morphological bit planes in retinal blood vessel extraction. *Journal of Digital Imaging*, 26(2):274–286, 2013.

[63] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, and S. A. Barman. Blood vessel segmentation methodologies in retinal images–a survey. *Computer Methods and Programs in Biomedicine*, 108(1):407–433, 2012.

[64] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, and S. A. Barman. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Transactions on Biomedical Engineering*, 59(9):2538–2548, 2012.

[65] M. M. Fraz, A. R. Rudnicka, C. G. Owen, and S. A. Barman. Delineation of blood vessels in pediatric retinal images using decision trees-based ensemble classification. *International Journal of Computer Assisted Radiology and Surgery*, 9(5):795–811, 2014.

[66] H. Fu, Y. Xu, S. Lin, D. W. K. Wong, and J. Liu. DeepVessel: Retinal vessel segmentation via deep learning and conditional random field. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 132–139. Springer, 2016.

[67] H. Fu, Y. Xu, D. W. K. Wong, and J. Liu. Retinal vessel segmentation via deep learning network and fully-connected conditional random fields. In *IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pp. 698–701. IEEE, 2016.

[68] G. Gardner, D. Keating, T. Williamson, and A. Elliott. Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool. *British Journal of Ophthalmology*, 80(11):940–944, 1996.

[69] R. Gargeya and T. Leng. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 2017.

[70] L. Giancardo, T. P. Karnowski, K. W. Tobin, F. Meriaudeau, and E. Chaum. Validation of microaneurysm-based diabetic retinopathy screening across retina fundus datasets. In *IEEE 26th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 125–130. IEEE, 2013.

[71] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT Press, 2016.

[72] J. Grauslund, A. Green, R. Kawasaki, L. Hodgson, A. K. Sjølie, and T. Y. Wong. Retinal vascular fractals and microvascular and macrovascular complications in type 1 diabetes. *Ophthalmology*, 117(7):1400–1405, 2010.

[73] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016.

[74] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[75] M. S. Haleem, L. Han, J. van Hemert, and B. Li. Automatic extraction of retinal features from colour retinal images for glaucoma diagnosis: A review. *Computerized Medical Imaging and Graphics*, 37(7):581–596, 2013.

[76] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[77] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*, volume 1. Springer, 2009.

[78] H. He, E. Garcia, et al. Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284, 2009.

[79] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán. Multiscale conditional random fields for image labeling. In *Proceedings of the 2004 IEEE Computer*

Society Conference on Computer Vision and Pattern Recognition (CVPR), volume 2, pp. II–695. IEEE, 2004.

[80] M. Helmstaedter. Cellular-resolution connectomics: challenges of dense neural circuit reconstruction. *Nature Methods*, 10(6):501–507, 2013.

[81] E. M. Hoffmann, L. M. Zangwill, J. G. Crowston, and R. N. Weinreb. Optic disk size and glaucoma. *Survey of Ophthalmology*, 52(1):32–49, 2007.

[82] S. Holm, G. Russell, V. Nourrit, and N. McLoughlin. DR HAGIS: a fundus image database for the automatic extraction of retinal surface vessels from diabetic patients. *Journal of Medical Imaging*, 4(1):014503–014503, 2017.

[83] A. Hoover, V. Kouznetsova, and M. Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical Imaging*, 19(3):203–210, 2000.

[84] F. Huang, B. Dashtbozorg, J. Zhang, E. Bekkers, S. Abbasi-Sureshjani, T. T. Berendschot, and B. M. ter Haar Romeny. Reliability of using retinal vascular fractal dimension as a biomarker in the diabetic retinopathy detection. *Journal of Ophthalmology*, 2016, 2016.

[85] M. K. Ikram, S. de Voogd, R. C. Wolfs, A. Hofman, M. M. Breteler, L. D. Hubbard, and P. T. de Jong. Retinal vessel diameters and incident open-angle glaucoma and optic disc changes: the Rotterdam study. *Investigative Ophthalmology & Visual Science*, 46(4):1182–1187, 2005.

[86] M. K. Ikram, Y. T. Ong, C. Y. Cheung, and T. Y. Wong. Retinal vascular caliber measurements: clinical significance, current knowledge and future perspectives. *Ophthalmologica*, 229(3):125–136, 2012.

[87] M. Jaderberg, A. Vedaldi, and A. Zisserman. Speeding up convolutional neural networks with low rank expansions. *arXiv preprint arXiv:1405.3866*, 2014.

[88] T. Joachims et al. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1):27–59, 2009.

[89] J. B. Jonas, W. M. Budde, and S. Panda-Jonas. Ophthalmoscopic evaluation of the optic nerve head. *Survey of Ophthalmology*, 43(4):293–320, 1999.

[90] J. B. Jonas, X. N. Nguyen, and G. Naumann. Parapapillary retinal vessel diameter in normal and glaucoma eyes. I. morphometric data. *Investigative Ophthalmology & Visual Science*, 30(7):1599–1603, 1989.

[91] G. D. Joshi, J. Sivaswamy, and S. Krishnadas. Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment. *IEEE Transactions on Medical Imaging*, 30(6):1192–1205, 2011.

[92] S. Kadoury, N. Abi-Jaoudeh, and P. A. Valdes. Higher-order CRF tumor segmentation with discriminant manifold potentials. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*, pp. 719–726. Springer, 2013.

[93] S. Kadoury et al. Metastatic liver tumour segmentation from discriminant Grassmannian manifolds. *Physics in Medicine and Biology*, 60(16):6459, 2015.

[94] Kaggle. Diabetic Retinopathy Detection. `https://www.kaggle.com/c/diabetic-retinopathy-detection`, 2015. [Online; accessed 10-April-2013].

[95] G. B. Kande, P. V. Subbaiah, and T. S. Savithri. Unsupervised fuzzy based vessel segmentation in pathological digital fundus images. *Journal of Medical Systems*, 34(5):849–858, 2010.

[96] P. L. Kaufman, F. H. Adler, L. A. Levin, and A. Alm. *Adler's Physiology of the Eye*. Elsevier Health Sciences, 2011.

[97] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kälviäinen, and J. Pietilä. The DIARETDB1 diabetic retinopathy database and evaluation protocol. In *Proceedings of the British Machine Vision Conference*. 2007.

[98] V. Koh, C. Y.-l. Cheung, Y. Zheng, T. Y. Wong, W. Wong, and T. Aung. Relationship of retinal vascular tortuosity with the neuroretinal rim: The Singapore

Malay eye study. *Investigative Ophthalmology & Visual Science*, 51(7):3736–3741, 2010.

[99] T. Köhler, R. Bock, J. Hornegger, and G. Michelson. Computer-aided diagnostics and pattern recognition: Automated glaucoma detection. In *Teleophthalmology in Preventive Medicine*, pp. 93–104. Springer, 2015.

[100] R. Kolar, R. P. Tornow, R. Laemmer, J. Odstrcilik, M. A. Mayer, J. Gazarek, J. Jan, T. Kubena, and P. Cernosek. Analysis of visual appearance of retinal nerve fibers in high resolution fundus images: a study on normal subjects. *Computational and Mathematical Methods in Medicine*, 2013, 2013.

[101] N. Komodakis et al. MRF energy minimization and beyond via dual decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):531–552, 2011.

[102] P. Krähenbühl and V. Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *Advances in Neural Information Processing Systems*, pp. 109–117. 2012.

[103] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097–1105. 2012.

[104] S. Kumar and M. Hebert. Discriminative random fields. *International Journal of Computer Vision*, 68(2):179–201, 2006. ISSN 0920-5691. doi: 10.1007/s11263-006-7007-9.

[105] J. D. Lafferty et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289. Morgan Kaufmann Publishers Inc., 2001.

[106] J. L. Leasher, R. R. Bourne, S. R. Flaxman, J. B. Jonas, J. Keeffe, K. Naidoo, K. Pesudovs, H. Price, R. A. White, T. Y. Wong, et al. Global estimates on the number of people blind or visually impaired by diabetic retinopathy: A meta-analysis from 1990 to 2010. *Diabetes Care*, 39(9):1643–1649, 2016.

[107] J. L. Leasher, V. Lansingh, S. R. Flaxman, J. B. Jonas, J. Keeffe, K. Naidoo, K. Pesudovs, H. Price, J. C. Silva, R. A. White, et al. Prevalence and causes of vision loss in Latin America and the Caribbean: 1990–2010. *British Journal of Ophthalmology*, pp. bjophthalmol–2013, 2014.

[108] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[109] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[110] J. Lee, B. C. Y. Zee, and Q. Li. Detection of neovascularization based on fractal and texture analysis with interaction effects in diabetic retinopathy. *PLOS ONE*, 8(12):e75699, 2013.

[111] M. C. Leske, A. Connell, S.-Y. Wu, L. G. Hyman, and A. P. Schachat. Risk factors for open-angle glaucoma: the Barbados Eye Study. *Archives of Ophthalmology*, 113(7):918–924, 1995.

[112] K. Li, X. Wu, D. Z. Chen, and M. Sonka. Optimal surface segmentation in volumetric images-a graph-theoretic approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):119–134, 2006.

[113] S. Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer, 3rd ed$^{\underline{\text{ón}}}$., 2009.

[114] D. Y. Lin, M. S. Blumenkranz, R. J. Brothers, D. M. Grosvenor, and T. D. D. S. Group. The sensitivity and specificity of single-field nonmydriatic monochromatic digital fundus photography with remote image interpretation for diabetic retinopathy screening: a comparison with ophthalmoscopy and standardized mydriatic color photography. *American Journal of Ophthalmology*, 134(2):204–213, 2002.

[115] L. Lo Vercio, M. del Fresno, and I. Larrabide. Detection of morphological structures for vessel wall segmentation in IVUS using random forests. In *12th*

International Symposium on Medical Information Processing and Analysis, pp. 1016012–1016012. International Society for Optics and Photonics, 2017.

[116] L. Lo Vercio, J. I. Orlando, M. del Fresno, and I. Larrabide. Assessment of image features for vessel wall segmentation in intravascular ultrasound images. *International Journal of Computer Assisted Radiology and Surgery*, 11(8):1397–1407, 2016.

[117] C.-K. Lu, T. B. Tang, A. Laude, B. Dhillon, and A. F. Murray. Parapapillary atrophy and optic disc region assessment (PANDORA): retinal imaging tool for assessment of the optic disc and parapapillary atrophy. *Journal of Biomedical Optics*, 17(10):1060101–1060108, 2012.

[118] C. A. Lupascu, D. Tegolo, and E. Trucco. FABC: retinal vessel segmentation using AdaBoost. *IEEE Transactions on Information Technology in Biomedicine*, 14(5):1267–1274, 2010.

[119] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool. Deep retinal image understanding. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 140–148. Springer, 2016.

[120] H. Manterola, M. del Fresno, and I. Larrabide. An analysis of mechanical and computational properties for noninvasive vascular elastography. In *11th International Symposium on Medical Information Processing and Analysis (SIPAIM 2015)*, pp. 968113–968113. International Society for Optics and Photonics, 2015.

[121] D. Marín, A. Aquino, M. E. Gegúndez-Arias, and J. M. Bravo. A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features. *IEEE Transactions on Medical Imaging*, 30(1):146–158, 2011.

[122] M. E. Martinez-Perez, A. Highes, A. V. Stanton, S. A. Thorn, N. Chapman, A. A. Bharath, and K. H. Parker. Retinal vascular tree morphology: a semi-automatic quantification. *IEEE Transactions on Biomedical Engineering*, 49(8):912–917, 2002.

[123] M. E. Martinez-Perez, A. D. Hughes, S. A. Thom, A. A. Bharath, and K. H. Parker. Segmentation of blood vessels from red-free and fluorescein retinal images. *Medical Image Analysis*, 11(1):47–61, 2007.

[124] J. Meier, R. Bock, G. Michelson, L. G. Nyúl, and J. Hornegger. Effects of preprocessing eye fundus images on appearance based glaucoma classification. In *Computer Analysis of Images and Patterns*, pp. 165–172. Springer, 2007.

[125] J. Meier, R. Bock, L. G. Nyúl, and G. Michelson. Eye fundus image processing system for automated glaucoma classification. In *Proceedings of the 52nd Internationales Wissenschaftliches Kolloquium*. Technische Universität Ilmenau, 2007.

[126] A. M. Mendonca and A. Campilho. Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction. *IEEE Transactions on Medical Imaging*, 25(9), 2006.

[127] G. Michelson et al. *Teleophthalmology in Preventive Medicine*. Springer, 2015.

[128] M. S. Miri and A. Mahloojifar. Retinal image analysis using curvelet transform and multistructure elements morphology by reconstruction. *IEEE Transactions on Biomedical Engineering*, 58(5):1183–1192, 2011.

[129] P. Mitchell, H. Leung, J. J. Wang, E. Rochtchina, A. J. Lee, T. Y. Wong, and R. Klein. Retinal vessel diameter and open-angle glaucoma: the Blue Mountains Eye Study. *Ophthalmology*, 112(2):245–250, 2005.

[130] S. Mohammad and D. T. Morris. Texture analysis for glaucoma classification. In *International Conference on BioSignal Analysis, Processing and Systems (ICBAPS)*, pp. 98–103. IEEE, 2015.

[131] M. R. K. Mookiah, U. R. Acharya, C. K. Chua, C. M. Lim, E. Ng, and A. Laude. Computer-aided diagnosis of diabetic retinopathy: A review. *Computers in Biology and Medicine*, 43(12):2136–2155, 2013.

[132] M. Mozaffarieh, M. C. Grieshaber, and J. Flammer. Oxygen and blood flow: players in the pathogenesis of glaucoma. *Molecular Vision*, 14:224–233, 2008.

[133] C. Muramatsu, Y. Hayashi, A. Sawada, Y. Hatanaka, T. Hara, T. Yamamoto, and H. Fujita. Detection of retinal nerve fiber layer defects on retinal fundus images for early diagnosis of glaucoma. *Journal of Biomedical Optics*, 15(1):016021–016021, 2010.

[134] J. Nandy, W. Hsu, and M. L. Lee. An incremental feature extraction framework for referable diabetic retinopathy detection. In *IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 908–912. IEEE, 2016.

[135] A. Y. Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Proceedings of the International Conference on Machine Learning*, p. 78. 2004.

[136] H. V. Nguyen, G. S. W. Tan, R. J. Tapp, S. Mital, D. S. W. Ting, H. T. Wong, C. S. Tan, A. Laude, E. S. Tai, N. C. Tan, et al. Cost-effectiveness of a national telemedicine diabetic retinopathy screening program in singapore. *Ophthalmology*, 123(12):2571–2580, 2016.

[137] T. T. Nguyen, J. J. Wang, A. R. Sharrett, F. A. Islam, R. Klein, B. E. Klein, M. F. Cotch, and T. Y. Wong. Relationship of retinal vascular caliber with diabetes and retinopathy. *Diabetes Care*, 31(3):544–549, 2008.

[138] U. T. Nguyen et al. An effective retinal blood vessel segmentation method using multi-scale line detection. *Pattern Recognition*, 46(3):703–715, 2013.

[139] M. Niemeijer, M. D. Abramoff, and B. Van Ginneken. Information fusion for diabetic retinopathy cad in digital color fundus photographs. *IEEE Transactions on Medical Imaging*, 28(5):775–785, 2009.

[140] M. Niemeijer, M. K. Garvin, B. van Ginneken, M. Sonka, and M. D. Abramoff. Vessel segmentation in 3D spectral OCT scans of the retina. In *Medical Imaging*, pp. 69141R–69141R. International Society for Optics and Photonics, 2008.

[141] M. Niemeijer, J. Staal, B. van Ginneken, M. Loog, M. D. Abramoff, et al. Comparative study of retinal vessel segmentation methods on a new publicly

available database. In *Medical Imaging 2004*, pp. 648–656. International Society for Optics and Photonics, 2004.

[142] M. Niemeijer, B. Van Ginneken, M. J. Cree, A. Mizutani, G. Quellec, C. I. Sánchez, B. Zhang, R. Hornero, M. Lamard, C. Muramatsu, et al. Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. *IEEE Transactions on Medical Imaging*, 29(1):185–195, 2010.

[143] M. Niemeijer, B. Van Ginneken, J. Staal, M. S. Suttorp-Schulten, and M. D. Abràmoff. Automatic detection of red lesions in digital color fundus photographs. *IEEE Transaction on Medical Imaging*, 24(5):584–592, 2005.

[144] J. Odstrčilík, J. Jan, J. Gazárek, and R. Kolář. Improvement of vessel segmentation by matched filtering in colour retinal images. In *World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany*, pp. 327–330. Springer, 2009.

[145] J. Odstrčilík, R. Kolar, A. Budai, J. Hornegger, J. Jan, J. Gazarek, T. Kubena, P. Cernosek, O. Svoboda, and E. Angelopoulou. Retinal vessel segmentation by improved matched filtering: evaluation on a new high-resolution fundus image database. *IET Image Processing*, 7(4):373–383, 2013.

[146] J. Odstrcilik, R. Kolar, J. Jan, J. Gazarek, Z. Kuna, and M. Vodakova. Analysis of retinal nerve fiber layer via markov random fields in color fundus images. In *19th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 504–507. IEEE, 2012.

[147] J. Odstrcilik, R. Kolar, R.-P. Tornow, J. Jan, A. Budai, M. Mayer, M. Vodakova, R. Laemmer, M. Lamos, Z. Kuna, et al. Thickness related textural properties of retinal nerve fiber layer in color fundus images. *Computerized Medical Imaging and Graphics*, 38(6):508–516, 2014.

[148] V. Optical. Volk Optical - Ophthalmic Imaging Designer & Manufacturer — Lenses — Surgical — Cameras. `https://volk.com/`. Accessed: 2017-03-23.

[149] W. H. Organization et al. Global data on visual impairments 2010. *URL: http://www. who. int/blindness/GLOBALDATAFINALforweb. pdf [accessed 2013-02-28][WebCite Cache]*, 2012.

[150] J. I. Orlando, E. Prokofyeva, M. del Fresno, and M. B. Blaschko. Convolutional neural network transfer for automated glaucoma identification. volume 10160, pp. 101600U–101600U–10. 2017. doi: 10.1117/12.2255740. URL `http://dx.doi.org/10.1117/12.2255740`.

[151] J. I. Orlando and M. B. Blaschko. Learning fully-connected CRFs for blood vessel segmentation in retinal images. In P. Golland, C. Barillot, J. Hornegger, and R. Howe, eds., *MICCAI 2014, LNCS*, volume 8149, pp. 634–641. Springer, 2014.

[152] J. I. Orlando and M. del Fresno. Reviewing preprocessing and feature extraction techniques for retinal blood vessel segmentation in fundus images. *Mecánica Computacional*, XXXIII(42):2729–2743, 2014.

[153] J. I. Orlando, H. L. Manterola, E. Ferrante, and F. Ariel. Arabidopsis roots segmentation based on morphological operations and CRFs. *arXiv preprint arXiv:1704.07793*, 2014.

[154] J. I. Orlando, E. Prokofyeva, and M. B. Blaschko. A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images. *IEEE Transactions on Biomedical Engineering*, 64(1):16–27, 2017.

[155] J. I. Orlando, E. Prokofyeva, M. del Fresno, and M. B. Blaschko. Learning to detect red lesions in fundus photographs: An ensemble approach based on deep learning. *arXiv preprint arXiv:1706.03008*, 2017.

[156] M. A. Palomera-Pérez, M. E. Martinez-Perez, H. Benítez-Pérez, and J. L. Ortega-Arjona. Parallel multiscale feature extraction and region growing: application in retinal blood vessel detection. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):500–506, 2010.

[157] N. Panwar, P. Huang, J. Lee, P. A. Keane, T. S. Chuan, A. Richhariya, S. Teoh, T. H. Lim, and R. Agrawal. Fundus photography in the 21st centurya review of recent technological advances and their implications for worldwide healthcare. *Telemedicine and e-Health*, 22(3):198–208, 2016.

[158] N. Patton, T. M. Aslam, T. MacGillivray, I. J. Deary, B. Dhillon, R. H. Eikelboom, K. Yogesan, and I. J. Constable. Retinal image analysis: concepts, applications and potential. *Progress in Retinal and Eye Research*, 25(1):99–127, 2006.

[159] A. Perez-Rovira, T. MacGillivray, E. Trucco, K. Chin, K. Zutis, C. Lupascu, D. Tegolo, A. Giachetti, P. Wilson, A. Doney, et al. VAMPIRE: vessel assessment and measurement platform for images of the retina. In *Engineering in medicine and biology society, EMBC, 2011 annual international conference of the IEEE*, pp. 3391–3394. IEEE, 2011.

[160] A. Perez-Rovira, K. Zutis, J. Hubschman, and E. Trucco. Improving vessel segmentation in ultra-wide field-of-view retinal fluorescein angiograms. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2614–2617. IEEE, 2011.

[161] R. Pires, S. Avila, H. Jelinek, J. Wainer, E. Valle, and A. Rocha. Beyond lesion-based diabetic retinopathy: a direct approach for referral. *IEEE Journal of Biomedical and Health Informatics*, 2015.

[162] R. Pires, H. F. Jelinek, J. Wainer, S. Goldenstein, E. Valle, and A. Rocha. Assessing the need for referral in automatic diabetic retinopathy detection. *IEEE Transactions on Biomedical Engineering*, 60(12):3391–3398, 2013.

[163] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng. Convolutional neural networks for diabetic retinopathy. *Procedia Computer Science*, 90:200–205, 2016.

[164] E. Prokofyeva and E. Zrenner. Epidemiology of major eye diseases leading to blindness in Europe: A literature review. *Ophthalmic Research*, 47(4):171–188, 2012. doi: 10.1159/000329603.

[165] G. Quellec, K. Charrière, Y. Boudi, B. Cochener, and M. Lamard. Deep image mining for diabetic retinopathy screening. *arXiv preprint arXiv:1610.07086*, 2016.

[166] G. Quellec, M. Lamard, A. Erginay, A. Chabouis, P. Massin, B. Cochener, and G. Cazuguel. Automatic detection of referral patients due to retinal pathologies through data mining. *Medical Image Analysis*, 29:47–64, 2016.

[167] R. Ranjan. Fundus Fluoroscein Angiography. `https://es.slideshare.net/rashmiranjan589/fundus-fluoroscein-angiography/7?smtNoRedir=1`. Accessed: 2017-03-23.

[168] S. Ravishankar, A. Jain, and A. Mittal. Automated feature extraction for early detection of diabetic retinopathy in fundus images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 210–217. IEEE, 2009.

[169] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 512–519. IEEE, 2014.

[170] G. Richard, G. Soubrane, and L. Yannuzzi. Fluorescein angiography: textbook and atlas (2nd rev. and expanded ed.). 1998.

[171] R. Roletschek. Wikimedia Commons: Fundus kamera by Ralf Roletschek / Wikipedia. `https://commons.wikimedia.org/wiki/File:2010-12-07-funduskamera-by-RalfR-02.jpg`. Accessed: 2017-03-23.

[172] S. Roychowdhury et al. Blood vessel segmentation of fundus images by major vessel extraction and sub-image classification. *IEEE Journal of Biomedical and Health Informatics*, PP(99):1–1, 2014.

[173] S. Roychowdhury, D. D. Koozekanani, and K. K. Parhi. Iterative vessel segmentation of fundus images. *IEEE Transactions on Biomedical Engineering*, 62(7):1738–1749, 2015.

[174] S. Roychowdhury, D. D. Koozekanani, and K. K. Parhi. Automated detection of neovascularization for proliferative diabetic retinopathy screening. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pp. 1300–1303. IEEE, 2016.

[175] A. Rubinstein, L. Gutierrez, A. Beratarrechea, and V. E. Irazola. Increased prevalence of diabetes in Argentina is due to easier health care access rather than to an actual increase in prevalence. *PLOS ONE*, 9(4):e92245, 2014.

[176] C. I. Sánchez, M. Niemeijer, A. V. Dumitrescu, M. S. Suttorp-Schulten, M. D. Abràmoff, and B. van Ginneken. Evaluation of a computer-aided diagnosis system for diabetic retinopathy screening on public data. *Investigative Ophthalmology & Visual Science*, 52(7):4866–4871, 2011.

[177] P. J. Savino and H. V. Danesh-Meyer. *Color Atlas and Synopsis of Clinical Ophthalmology–Wills Eye Institute–Neuro-Ophthalmology.* Lippincott Williams & Wilkins, 2012.

[178] B. Schölkopf. *Support Vector Learning.* PhD. Thesis, Oldenbourg Verlag, Munich, 1997.

[179] L. Seoud, J. Chelbi, and F. Cheriet. Automatic grading of diabetic retinopathy on a public database. In *Proceedings of the Ophthalmic Medical Image Analysis Second International Workshop, OMIA*. MICCAI, 2015.

[180] L. Seoud, T. Hurtut, J. Chelbi, F. Cheriet, and J. P. Langlois. Red lesion detection using dynamic shape features for diabetic retinopathy screening. *IEEE Transactions on Medical Imaging*, 35(4):1116–1126, 2016.

[181] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[182] M. Shakeri, S. Tsogkas, E. Ferrante, S. Lippe, S. Kadoury, N. Paragios, and I. Kokkinos. Sub-cortical brain structure segmentation using F-CNN's. In *IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pp. 269–272. IEEE, 2016.

[183] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 806–813. 2014.

[184] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016.

[185] H. Sidahmed, E. Prokofyeva, and M. B. Blaschko. Discovering predictors of mental health service utilization with k-support regularized logistic regression. *Information Sciences*, 329:937–949, 2016. doi: http://dx.doi.org/10.1016/j.ins.2015.03.069.

[186] D. A. Sim, P. A. Keane, A. Tufail, C. A. Egan, L. P. Aiello, and P. S. Silva. Automated retinal image analysis for diabetic retinopathy in telemedicine. *Current Diabetes Reports*, 15(3):1–9, 2015.

[187] J. Sivaswamy, S. Krishnadas, A. Chakravarty, G. Joshi, A. S. Tabish, et al. A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. *JSM Biomedical Imaging Data Papers*, 2(1), 2015.

[188] J. V. Soares et al. Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification. *IEEE Transactions on Medical Imaging*, 25(9), 2006.

[189] O. P. Society. Retinal OCT imaging. `http://www.opsweb.org/?page=RetinalOCT`, 2017. Accessed: 2017-03-27.

[190] A. Soltanipour et al. Vessel centerlines extraction from fundus fluorescein angiogram based on hessian analysis of directional curvelet subbands. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1070–1074. IEEE, 2013.

[191] M. Sonka and M. D. Abràmoff. Quantitative analysis of retinal OCT. 2016.

[192] G. Spaeth. Appearances of the optic disc in glaucoma: a pathogenetic classification. In *Trans Acad Ophthalmol Symposium on Glaucoma. CV Mosby, St. Louis*. 1981.

[193] J. Staal et al. Ridge based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4):501–509, 2004.

[194] Y.-C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C.-Y. Cheng. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*, 121(11):2081–2090, 2014.

[195] A. Trost, S. Lange, F. Schroedl, D. Bruckner, K. A. Motloch, B. Bogner, A. Kaser-Eichberger, C. Strohmaier, C. Runge, L. Aigner, et al. Brain and retinal pericytes: origin, function and role. *Frontiers in Cellular Neuroscience*, 10, 2016.

[196] E. Trucco, A. Ruggeri, T. Karnowski, L. Giancardo, E. Chaum, J. P. Hubschman, B. Al-Diri, C. Y. Cheung, D. Wong, M. Abramoff, et al. Validating retinal fundus image analysis algorithms: Issues and a proposal. *Investigative Ophthalmology & Visual Science*, 54(5):3546–3559, 2013.

[197] I. Tsochantaridis et al. Large margin methods for structured and interdependent output variables. In *Journal of Machine Learning Research*, pp. 1453–1484. 2005.

[198] L. Van Der Maaten. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

[199] B. van Ginneken, A. A. Setio, C. Jacobs, and F. Ciompi. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In *IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 286–289. IEEE, 2015.

[200] M. J. van Grinsven, B. van Ginneken, C. B. Hoyng, T. Theelen, and C. I. Sánchez. Fast convolutional neural network training using selective data sam-

pling: Application to hemorrhage detection in color fundus images. *IEEE Transactions on Medical Imaging*, 35(5):1273–1284, 2016.

[201] R. Vega et al. Retinal vessel extraction using lattice neural networks with dendritic processing. *Computers in Biology and Medicine*, 58:20–30, 2015.

[202] R. Venkataramani, S. Thiruvenkadam, P. Annangi, N. Babu, and V. Vaidya. Understanding the mechanisms of deep transfer learning for medical images. In *LABELS and DLMIA, at Medical Image Computing and Computer-Assisted Intervention-MICCAI*, volume 10008, p. 188. Springer, 2016.

[203] J. A. Vilensky, W. Robertson, and C. A. Suarez-Quian. *The Clinical Anatomy of the Cranial Nerves: The Nerves of" On Old Olympus Towering Top"*. John Wiley & Sons, 2015.

[204] M. Vlachos and E. Dermatas. Multi-scale retinal vessel segmentation using line tracking. *Computerized Medical Imaging and Graphics*, 34(3):213–227, 2010.

[205] H. H. Vo and A. Verma. New deep neural nets for fine-grained diabetic retinopathy recognition on hybrid color space. In *2016 IEEE International Symposium on Multimedia (ISM)*, pp. 209–215. IEEE, 2016.

[206] P. Vostatek, E. Claridge, H. Uusitalo, M. Hauta-Kasari, P. Fält, and L. Lensu. Performance comparison of publicly available retinal blood vessel segmentation methods. *Computerized Medical Imaging and Graphics*, 55:2–12, 2017.

[207] T. Walter, P. Massin, A. Erginay, R. Ordonez, C. Jeulin, and J.-C. Klein. Automatic detection of microaneurysms in color fundus images. *Medical Image Analysis*, 11(6):555–566, 2007.

[208] O. B. Walton, R. B. Garoon, C. Y. Weng, J. Gross, A. K. Young, K. A. Camero, H. Jin, P. E. Carvounis, R. E. Coffee, and Y. I. Chu. Evaluation of automated teleretinal screening program for diabetic retinopathy. *JAMA Ophthalmology*, 134(2):204–209, 2016.

[209] H. Wang, A. Cruz-Roa, A. Basavanhally, H. Gilmore, N. Shih, M. Feldman, J. Tomaszewski, F. Gonzalez, and A. Madabhushi. Mitosis detection in breast

cancer pathology images by combining handcrafted and convolutional neural network features. *Journal of Medical Imaging*, 1(3):034003–034003, 2014.

[210] L. Wang, A. Bhalerao, and R. Wilson. Analysis of retinal vasculature using a multiresolution hermite model. *IEEE Transactions on Medical Imaging*, 26(2):137–152, 2007.

[211] X. Wang, L. I. Mudie, M. Baskaran, C.-Y. Cheng, W. L. Alward, D. S. Friedman, and C. J. Brady. Crowdsourcing to evaluate fundus photographs for the presence of glaucoma. *Journal of Glaucoma*, 2017.

[212] Z. Wang and J. Yang. Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation. *arXiv preprint arXiv:1703.10757*, 2017.

[213] R. N. Weinreb, T. Aung, and F. A. Medeiros. The pathophysiology and treatment of glaucoma: a review. *JAMA*, 311(18):1901–1911, 2014.

[214] R. Welikala, J. Dehmeshki, A. Hoppe, V. Tah, S. Mann, T. H. Williamson, and S. Barman. Automated detection of proliferative diabetic retinopathy using a modified line operator and dual classification. *Computer Methods and Programs in Biomedicine*, 114(3):247–261, 2014.

[215] D. R. Whiting, L. Guariguata, C. Weil, and J. Shaw. IDF diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030. *Diabetes Research and Clinical Practice*, 94(3):311–321, 2011.

[216] K. Willekens, S. Bataillie, I. Sarens, S. Odent, L. Abegão Pinto, E. Vandewalle, K. Van Keer, and I. Stalmans. Funduscopic versus HRT III Confocal Scanner Vertical Cup-Disc Ratio Assessment in Normal Tension and Primary Open Angle Glaucoma (The Leuven Eye Study). *Ophthalmic Research*, 57(2):100–106, 2016.

[217] B. Wu, W. Zhu, F. Shi, S. Zhu, and X. Chen. Automatic detection of microaneurysms in retinal fundus images. *Computerized Medical Imaging and Graphics*, 55:106–112, 2017.

[218] R. Wu, C. Y.-L. Cheung, S. M. Saw, P. Mitchell, T. Aung, and T. Y. Wong. Retinal vascular geometry and glaucoma: the singapore malay eye study. *Ophthalmology*, 120(1):77–83, 2013.

[219] D. Xiao and Y. Kanagasingam. Screening of the retina in diabetes patients by morphological means. In *Teleophthalmology in Preventive Medicine*, pp. 15–26. Springer, 2015.

[220] L. Xu and S. Luo. A novel method for blood vessel detection from retinal images. *Biomedical Engineering Online*, 9(1):14, 2010.

[221] Y. Xu, L. Duan, D. W. K. Wong, T. Y. Wong, and J. Liu. Glaucoma detection by learning from multiple informatics domains. In E. Trucco, X. Chen, Garvin, M. K., J. J. Liu, and X. Y. Frank, eds., *Proceedings of the Ophthalmic Medical Image Analysis Second International Workshop, OMIA*. 2015.

[222] L. A. Yannuzzi, M. D. Ober, J. S. Slakter, R. F. Spaide, Y. L. Fisher, R. W. Flower, and R. Rosen. Ophthalmic fundus imaging: today and beyond. *American Journal of Ophthalmology*, 137(3):511–524, 2004.

[223] Y. Yin et al. Automatic segmentation and measurement of vasculature in retinal fundus images using probabilistic formulation. *Computational and Mathematical Methods in Medicine*, 2013, 2013.

[224] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pp. 3320–3328. 2014.

[225] X. You et al. Segmentation of retinal blood vessels using the radial projection and semi-supervised approach. *Pattern Recognition*, 44(10), 2011.

[226] F. Zana and J.-C. Klein. Segmentation of vessel-like patterns using mathematical morphology and curvature evaluation. *IEEE Transactions on Image Processing*, 10(7):1010–1019, 2001.

[227] B. Zhang, L. Zhang, L. Zhang, and F. Karray. Retinal vessel extraction by matched filter with first-order derivative of Gaussian. *Computers in Biology and Medicine*, 40(4):438–445, 2010.

[228] Y. Zhao et al. Automated vessel segmentation using infinite perimeter active contour model with hybrid region information with application to retina images. *IEEE Transactions on Medical Imaging*, 2015.

[229] L. Zheng, B. Gong, D. A. Hatala, and T. S. Kern. Retinal ischemia and reperfusion causes capillary degeneration: similarities to diabetes. *Investigative Ophthalmology & Visual Science*, 48(1):361–367, 2007.

[230] Y. Zheng, D. Liu, B. Georgescu, H. Nguyen, and D. Comaniciu. 3D deep learning for efficient and robust landmark detection in volumetric data. In *Medical Image Computing and Computer-Assisted Intervention-MICCAI*, pp. 565–572. 2015.